
# ADVANCED NUMERICAL TECHNIQUES IN PARTIAL DIFFERENTIAL EQUATIONS:
## NONSTANDARD BOUNDARY CONDITIONS AND REGULARIZATION OF INVERSE PROBLEMS

**VLADIMÍR VRÁBEL'**

*supervisor*: MARIÁN SLODIČKA

Thesis submitted to Ghent University in candidature for the academic degree of Doctor of Philosphy in Mathematics

Ghent University
Faculty of Engineering and Architecture
Department of Mathematical Analysis
Research Group for Numerical Analysis and Mathematical Modelling ($NaM^2$)

Academic year
2012-2013

# Acknowledgements

In the first place, I would like to thank my supervisor Marián Slodička that he has put the trust in me and offered me the opportunity to be his PhD student. I am grateful for his constant help during my doctoral studies and many valuable advices. At the same time, I would like to thank my colleague Valdemar Melicher with whom I could collaborate. I am happy that our long discussions and joint effort on his idea bore fruit, which is also a part of the thesis. My research has been supported by the BOF-grant number 01D00409 of Ghent University. I wish to express my appreciation for this indispensable financial support.

I am very grateful for meeting the right people in the right places. My hearty thanks go to my colleague Karel Van Bockstal who carefully read my thesis and suggested helpful corrections. I am also indebted to my teachers and professors for sharing the knowledge. Among them, special thanks go to Marián Fecko for teaching me mathematics from a different perspective and Ján Filo who recommended me Ghent when going on my Erasmus. I would like to really thank my colleagues and my friends here and in Slovakia for the nice times we had together.

I thank my parents and my family for all they have done for me. Last but not least, to my dear Ola, thank you for all the love and support you are giving me.

Vladimír Vrábeľ

# Contents

# Samenvatting

Dit doctoraatsproefschrift bestaat uit twee op zichzelf staande delen. In het eerste deel worden *randwaardeproblemen* met een klasse van *niet-standaard randvoorwaarden* bestudeerd. Het tweede deel behandelt een *continuatie methode* in *inverse problemen* en optimalisatie met *partiële differentiaalvergelijking* beperking. Aan het einde van het doctoraatsproefschrift bevindt zich een bijlage die de belangrijkste formules, definities, stellingen en de bibliografie bevat.

Partiële differentiaalvergelijkingen zijn vergelijkingen waarin, naast eventueel de onbekende functie zelf, ook partiële afgeleiden van die functie voorkomen. Zij maken een onontbeerlijk deel uit van de wiskundige modellering van diverse natuurverschijnselen. Modellen met partiële differentiaalvergelijkingen worden in de numerieke wiskunde vaak voorgesteld als randwaardeproblemen. Een randwaardeprobleem is een (partiële) differentiaalvergelijking op een begrensd domein samen met een reeks bijkomende beperkingen, zogenaamde randvoorwaarden. Een oplossing van een randwaardeprobleem is een oplossing van de differentiaalvergelijking die aan de randvoorwaarden moet voldoen. Er bestaan drie soorten standaard randvoorwaarden (randcondities). De Dirichlet randconditie bepaalt de waarden die de oplossing moet hebben aan de rand van het domein. De Neumann randconditie bepaalt de waarden die de normale afgeleide van de oplossing moet hebben aan de rand van het domein. De derde soort, de Robin randconditie, is een lineaire combinatie van beide. Tijdsafhankelijke randwaardeproblemen vereisen bovendien een beginvoorwaarde. Naast de vermelde standaard randvoorwaarden is het soms geschikt om gebruik te maken van niet-standaard randvoorwaarden. Zij stellen vaak een interessante wisselwerking voor tussen de nauwkeurigheid van een fysisch model en de computationele kosten van een numerieke model. De niet-standaard randvoorwaarden komen voor in diverse situaties en daarom beperken we in dit werk niet tot één natuurkundige model.

## Deel I

Het eerste deel wordt toegewijd aan de analyse van verschillende parabolische randwaardeproblemen met een klasse van niet-lineaire niet-standaard randvoorwaarden. We

bekomen de existentie en uniciteit van zwakke oplossingen van die randwaardeproblemen in geschikte functieruimten. De analyse steunt op de Rothe methode en de theorie van monotone operatoren, zie de referenties [71] en [126]. We discretiseren de randwaardeproblemen in tijd door gebruik maken van de achterwaartse Euler methode. Dan passen we de theorie van monotone operatoren toe om aan te tonen dat het tijdsgediscretiseerde probleem op elke tijdstap uniek oplosbaar is. Op die manier construeren we de Rothe functierij wiens functies stukgewijs gedefinieerd zijn. Vervolgens wordt de convergentie van dit benaderingsschema naar de oplossing van het origineel probleem aangetoond en de hierbij horende foutenschattingen voor tijdsdiscretisatie en ruimtediscretisatie worden opgesteld.

Tot slot worden numerieke experimenten uitgevoerd voor elk randwaardeprobleem om de performantie van de methoden en de theoretische resultaten te testen. De numerieke implementatie volgt nauwkeurig de theoretische analyse: we gebruiken de achterwaartse Euler methode voor de tijdsdiscretisatie en de eindige elementen methode voor de ruimtediscretisatie [31]. Al de numerieke berekeningen zijn uitgevoerd met behulp van het *softwarepakket FEniCS* [78].

Het eerste deel wordt verdeeld in drie hoofdstukken. In Hoofdstuk 2 beschouwen we een randwaardeprobleem met niet-lineaire dynamische randconditie. De motivatie voor dit probleem komt van de modellering van regenwaterinfiltratie in de bodem. De standaard Neumann randconditie, die enkel neerslag intensiteit op het aardoppervlak beschrijft, houdt geen rekening met een in de praktijk mogelijk saturatie effect. In dit geval sijpelt het water niet in de grond met de vorming van waterplassen tot gevolg. Een manier om dit probleem te behandelen en het model te verbeteren is een dynamische randconditie die een tijdsafgeleide bevat. We bewijzen dat het correspondende wiskundige model goed gesteld is. Dit hoofdstuk dient ook als introductie in de Rothe methode. Er worden enkel foutenschattingen voor de tijdsdiscretisatie gemaakt. De resultaten zijn voorgesteld op de conferentie ACOMEN 2011, the Fifth International Conference on Advanced Computational Methods in Engineering, in Liege en vervolgens gepublicieerd in Journal of Computational and Applied Mathematics, [118].

In Hoofdstuk 3 bestuderen we een niet-lineaire warmtediffusie vergelijking met een diffusie-achtige randconditie. Die bevat onder andere de Laplace-Beltrami operator, die werkt in de tangentiële richting aan de rand en die modeleert de diffusie langs de rand. In het begin van het hoofdstuk plaatsen we de randconditie in kwestie in het bredere kader van standaard en niet-standaard randvoorwaarden voor parabolische randwaardeproblemen. Een eenvoudig voorbeeld van dit soort randwaardeproblemen is de lineaire warmtevergelijking voor de functie $u$ op het domein $\Omega$ met rand $\partial\Omega$, die vergezeld is met de volgende algemene randconditie

$$\begin{aligned} &\partial_t u = -\alpha\Delta u \qquad \text{in } \Omega \text{ and } t > 0, \\ &\partial_t u = -\alpha\nabla u \cdot \boldsymbol{n} - \beta(u - u_{\text{out}}) + \Delta_\Gamma u \qquad \text{on } \partial\Omega \text{ and } t > 0, \end{aligned}$$

waar $\alpha$ en $\beta$ positieve constanten zijn en $u_{\text{out}}$ bekend is. De bovenstaande randcondi-

tie stelt dat de verandering in tijd van de temperatuur $u$ op de rand afhankelijk is van de warmteflux over de rand, het temperatuurverschil tussen de rand en het buitengebied, en van de warmtediffusie stroming langs de rand, die vertegenwoordigd is door de Laplace-Beltrami operator $\Delta_\Gamma u$ (vergelijk met [54]). De resultaten van dit hoofdstuk zijn voorgelegd aan het vaktijdschrift Applied Mathematics and Computations.

Hoofdstuk 4 begint met een kort overzicht van Maxwellvergelijkingen en aanvullende randvoorwaarden. Het hoofdgedeelte van dit hoofdstuk wordt ontleed aan het artikel [117], dat gepubliceerd is in Journal of Mathematical Analysis and Applications. We beschouwen de quasi-statische Maxwellvergelijkingen in het tijdsdomein met een niet-lineaire impedantie-achtige randconditie. De klassieke impedantie randvoorwaarde is een benaderende randvoorwaarde om het skineffect in geleiders te beschrijven. We merken op dat de randvoorwaarde in dit hoofdstuk kan gezien worden als een wiskundige analogie van die in Hoofdstuk 2 maar met de rotor-operator.

# Deel II

Het tweede deel wordt gebaseerd op een gezamenlijk artikel met Valdemar Melicher als medeauteur dat is voorgelegd aan het tijdschrift Inverse Problems. In het artikel stellen we een continuatie methode voor Tikhonov regularisatie voor met een speciale focus op inverse problemen, waar het direct probleem een partiële differentiaalvergelijking is.

In grote lijnen, bij inverse problemen moet men onbekende parameters van een model herwinnen aan de hand van bijkomende metingen. Deze problemen zijn inherent gedreven door toepassingen en doen zich voor in een grote verscheidenheid van praktische situaties. Een invers probleem veronderstelt een *direct probleem*, dat de parameters koppelt met die metingen. Het direct probleem is goed gesteld in de zin van Hadamard: er bestaat een unieke oplossing van het probleem, die continu afhangt van de data. Het kan bijvoorbeeld een partiële differentiaalvergelijking zijn, die een fysisch systeem modelleert. Een bijbehorende invers probleem is dan typisch een identificatie van een parameter van het onderliggende fysische model op basis van de bijkomende metingen van het systeem (het zogenaamde *parameter identificatieprobleem*). Invers problemen voldoen vaak niet aan de alle criteria om een goed gesteld probleem te zijn; ze zijn *slecht gestelde problemen*, wat ze zeer uitdagend maakt op te lossen.

Een typisch voorbeeld van een slecht gesteld probleem is de operatorvergelijking

$$F(u) = v, \tag{1}$$

waar $F : \mathcal{D}(F) \subset U \to V$ een *compleet continue* operator is tussen de Banachruimten $U$ en $V$.

Numerieke methoden die kunnen omgaan met het slecht gesteld karakter van inverse problemen zijn de zogenaamde *regularisatiemethoden*. De regularisatie is de benadering van een slecht gesteld probleem door een familie van benaderende goed gestelde problemen. Men neemt aan dat alleen de verstoorde data $v^\delta$ van de exacte data $v$ beschikbaar is

met een verstoringsniveau $\delta$. De meest gebruikte methode voor regularisatie van slecht gestelde problemen is de *Tikhonov regularisatie*. Met deze methode kan men zoeken naar de best benaderende oplossing van (1), die de volgende *Tikhonov functionaal* minimaliseert

$$\mathcal{T}_\alpha(u) = \left\|F(u) - v^\delta\right\|^2 + \alpha\mathcal{R}_U(u), \qquad \alpha > 0, \tag{2}$$

waar $\mathcal{R}_U(u)$ een regularisatie term is. In de praktijk worden *gradient methoden* vaak gebruikt om de minimiser $u_\alpha^\delta$ van $\mathcal{T}_\alpha$ te vinden. Voor meer info over inverse problemen zie de klassieke referentie [42].

De regularisatie term heeft twee hoofdtaken: stabilisatie van het slecht gesteld probleem en de a priori kennis over de oplossing invoeren. Het belangrijkste idee van het artikel bestaat in de onafhankelijke behandeling van deze twee rollen. We stabiliseren eerst het probleem in een grotere ruimte $W$ en dan transformeren we dit probleem via de continuatie methode in het oorspronkelijke probleem, dat is gesteld op de deelruimte $U \subset W$, waar we de gewenste oplossing verwachten.

Het tweede deel wordt gesplitst in twee hoofdstukken. In Hoofdstuk 5 stellen we de continuatie methode voor Tikhonov regularisatie in het algemeen voor. In plaats van de functionaal (2) te beschouwen, maken we gebruik van de volgende uitbreiding van de Tikhonov functionaal

$$\mathcal{T}_{\alpha,\beta}(u, w, \lambda) = \left\|F(z) - v^\delta\right\|^2 + \lambda\alpha\mathcal{R}_U(u) + (1-\lambda)\beta\mathcal{R}_W(w), \tag{3}$$

met $\lambda \in [0, 1]$ en

$$z = \lambda u + (1-\lambda)w.$$

We analyseren de goed-gesteldheid, de stabiliteit en de convergentie van deze functionaal. Het belangrijkste resultaat (Stelling 5.1) houdt in dat de voorgestelde continuatie methode lokaal correct is ten opzichte van de oplossing van het oorspronkelijke probleem.

Hoofdstuk 6 behandelt deze continuatie methode in de context van stuksgewijze constante parameter identificatieproblemen. Ze kunnen vaak worden gezien als geometrie identificatieproblemen. Een kort overzicht over meetkundige identificatie wordt gegeven met nadruk op de methoden die samen topologie- en vorm- sensitiviteitsconcepten combineren. Daarna stellen we de topology-to-shape continuatie methode (TSCM) voor. Deze benadering kan worden gebruikt om de lokale convergentie van gradient methoden in topologie-tot-vorm optimalisatie te vermijden. Deze aanpak is een veelbelovende kandidaat voor een kader dat zowel topologie als vorm sensitiviteiten verenigt.

In Sectie 6.2 tonen we illustratieve resultaten voor magnetische inductie tomografie aan, die vele toepassingen, bijvoorbeeld in de biomedische imaging en het niet-destructief onderzoek van materialen. Het invers probleem hier is de reconstructie van de stuksgewijze constante geleidbaarheid $\sigma$

$$\sigma_{PC} = \sigma_1\chi_D + \sigma_2\chi_{\Omega/D}, \quad \sigma_1, \sigma_2 \in \mathbb{R}, D \subset \Omega \tag{4}$$

in een lichaam (het domein $\Omega$) op basis van een eindig aantal Dirichlet-to-Neumann data, die overeenkomt met de impedantie afbeelding.

Sectie 6.2.5 bevat de implementatie van de TSCM en een paar numerieke experimenten voor magnetische inductie tomografie.

# Chapter 1

# Introduction

This thesis consists of two independent parts. The first one is devoted to *boundary value problems* with a certain class of *nonstandard boundary conditions*. The second part discusses a *continuation method in Tikhonov regularization* with a special focus on partial differential equation (PDE) constrained *inverse problems*.

## 1.1 Boundary value problems

*Partial differential equations* are equations involving unknown functions of two or more variables and some of their partial derivatives. The book [45] offers a comprehensive introduction to the modern theory of partial differential equations. They form an indispensable part of mathematical modeling of a wide variety of phenomena. Partial differential equations models in numerical mathematics are often stated as boundary value problems.

A boundary value problem (BVP) is a (partial) differential equation on a bounded domain together with a set of additional restraints, called the boundary conditions (BCs). A solution to a boundary value problem is a solution to the differential equation that satisfies the boundary conditions in the same time. There are three types of standard boundary conditions. The Dirichlet boundary condition specifies the value of a solution on the boundary of the domain. The Neumann boundary condition specifies the value of the normal derivative of a solution on the boundary of the domain. The third type, the Robin boundary condition, is a linear combination of both. Time-dependent boundary value problems moreover require an initial value condition.

In addition to the standard boundary conditions, it is sometimes convenient to make use of nonstandard boundary conditions. They often represent an interesting trade-off between the accuracy of a physical model and the computational cost of the numerical

one. Approximate boundary conditions in electromagnetism, for instance, are used to truncate the domain, which vastly reduces the costs. They are called by various names, such as absorbing, scattering, dissipative or reflecting boundary conditions, usually with respect to the property the authors want to stress.

Nonstandard boundary conditions can model thin layers, too. Take the impedance boundary condition in electromagnetism as a classical example. It replaces the material on one side of the boundary that is not a perfect conductor, but allows the electromagnetic field to penetrate only a small distance. In general case, one starts from a two-domain boundary value problem, where one domain is embedded into the second thin-layer domain. The main idea is to replace the latter one by a (nonstandard) boundary condition. One can model in this way complex phenomena on the boundary like surface currents in the case of electromagnetism or diffusion along the boundary in the context of anisotropic diffusion processes.

*Functional analysis* provides essential tools for studying boundary value problems [19]. One of the basic notions here is the one of a *weak (variational) formulation* of a boundary value problem. When dealing with nonlinear problems, the concept of a monotone operator plays a very important role. The best general reference for *monotone operator theory* and nonlinear functional analysis and its applications in general is the five volume work by Eberhard Zeidler, in particular [124, 125] and [126]. The reader is also invited to consult the appendix for the functional analysis used in this thesis.

The main analytical tool in the first part of the thesis is *Rothe's method* [71]. It provides a functional framework to establish existence and uniqueness of a solution to a time-dependent boundary value problem. It comes from the backward Euler method for solving time-dependent differential equations, where the original problem is replaced by a sequence of approximating time-discrete problems. The advantage of Rothe's method is that the solutions of those problems belong to the same functional spaces as the solution of the original problem. It is worth remarking that this method can be formulated using the semigroup theory [100].

In numerical experiments we use the finite element method to solve boundary value problems. The method is based on the weak formulation of the boundary value problem. An approximating solution is then searched in the spaces of piecewise-polynomial functions. The book [119] offers a systematic introduction to modern theory of partial differential equations and finite element methods. We also review the basic notions for this method in the appendix.

## 1.2 Inverse problems

Inverse problems are, roughly speaking, those where from measured data of a system one aims to recover the unknown model parameters of the system. In other words, the objective lies in converting observed measurements into information of interest about the system. Inverse problems are inherently driven by applications and they arise in

vast variety of practical situations. They are encountered in many branches of applied sciences such as biomedical engineering and imaging, geosciences, vulcanology, remote sensing, image and natural language processing and non-destructive material evaluation.

An inverse problem assumes a *direct (forward) problem*, which relates the model parameters to the measured data. The direct problem is well-posed in the sense of Hadamard. That is to say that,

(i) there exists a solution of the problem,

(ii) the solution is unique,

(iii) the solution depends continuously on the problem data.

A direct problem can be for instance a partial differential equation modeling a physical system. An associated inverse problem is then typically an identification of a physical parameter of the underlying physical model from measured observations of the system (so-called *parameter identification problem*). We refer the reader to [66] and [82] for more on inverse problems in partial differential equations.

Inverse problems often do not meet all the criteria of well-posedness; they are *ill-posed problems* what makes them very challenging to solve. Note that these criteria are of different degree of importance. The condition (i) appears not as restrictive, because it can be usually enforced by relaxing the notion of the solution. The violation of (ii) is considered to be much more serious. If a problem has several solutions, one either has to decide which one is of interest or one has to check the model for completeness and, if possible, feed in additional information. The violation of (iii) is symptomatic for inverse problems. Small *data perturbation* can significantly change the solution. This creates serious numerical problems, because, practically, one never knows exact data due to noise in measurements and errors in computations.

An useful illustration of an "inverse" and ill-posed problem is differentiation. Unlike integration, it does not have a smoothing property. Consider for instance a differentiable function $v \in C^1[0,1]$ and the associated sequence

$$v_n^\delta(x) := v(x) + \delta \sin \frac{nx}{\delta}, \qquad x \in [0,1],$$

where $n \in \mathbb{N}$ and $\delta \in (0,1)$. The function $v$ and $v_n^\delta$ represent here the exact and perturbed data, respectively. In the maximum norm we have

$$\max_{x\in[0,1]} |v - v_n^\delta| = \delta, \qquad \text{but} \qquad \max_{x\in[0,1]} |v' - (v_n^\delta)'| = n.$$

Therefore, the arbitrarily small *noise level* $\delta$ results in the arbitrarily large error in the derivative and so the differentiation in this context does not depend continuously on the data.

A typical example of an ill-posed problem is the operator equation

$$F(u) = v, \tag{1.1}$$

where $F : \mathcal{D}(F) \subset U \to V$ is a *completely continuous* operator between the Banach spaces $U$ and $V$. The operator $F$ is in the case of inverse problems associated with the forward problem. Recall that $F$ is *compact* if it maps bounded sets from $U$ into relatively compact sets in $V$. The operator $F$ is called completely continuous if it is compact and continuous. It follows from the definition that *linear compact* operators are always completely continuous. The well-posedness of the above problem means that $F : \mathcal{D}(F) \to \mathcal{R}(F)$ is injective in its range $\mathcal{R}(F) \subset V$ and the inverse operator $F^{-1} : \mathcal{R}(F) \to \mathcal{D}(F)$ is continuous. However, in the case that the domain $\mathcal{D}(F)$ is not finite dimensional, the problem (1.1) is ill-posed. Indeed, suppose for the sake of contradiction that $F^{-1}$ exists and it is continuous. Then from $I = F^{-1}F$ we see that the identity operator on $\mathcal{D}(F)$ is compact, since the composition of a continuous and a compact operator is compact. It is a well known fact that the identity operator in an infinite dimensional Banach space is not compact and thus the dimension of $\mathcal{D}(F)$ must be finite, which is the contradiction. [1]

Numerical methods that can cope with the ill-posed nature of inverse problems are the so-called *regularization methods*. The regularization is the approximation of an ill-posed problem by a family of neighbouring well-posed problems. The main goal is to find the best-approximate solution for the problem (1.1), where one assumes that only the noisy data $v^\delta$ of the exact data $v$ are available;

$$\left\| v - v^\delta \right\| \leq \delta$$

with $\delta$ being the noise level in some norm.

The most commonly used regularization method for ill-posed problems is *Tikhonov regularization* named after A. N. Tikhonov [113, 114]. Here, one can seek the best-approximate solution for (1.1) as a minimizer of a certain *Tikhonov functional*

$$\mathcal{T}_\alpha(u) = \left\| F(u) - v^\delta \right\|^2 + \alpha \left\| u \right\|^2, \qquad \alpha > 0. \tag{1.2}$$

It is in essence a trade-off between fitting the data and reducing a norm of the solution. The regularization term $\|u\|^2$ ("penalty term") stabilizes the problem and introduces the a priori knowledge about the solution. In general, it can be any proper convex functional. One of the standard ways how to choose the regularization parameter $\alpha$ is Morozov's *discrepancy principle*, which basically compares the residual error $\left\| F(u_\alpha^\delta) - v^\delta \right\|$ for the solution $u_\alpha^\delta$ of the minimization problem (1.2) with the noise level $\delta$. To numerically find the minimizer $u_\alpha^\delta$ of $\mathcal{T}_\alpha$, one often uses *gradient-based methods*. The approximative sequence $\{u_k\}$ for $u_\alpha^\delta$ is constructed as follows

$$u_k = u_{k-1} - \omega \mathcal{T}_\alpha'(u_{k-1}), \qquad k \in \mathbb{N},$$

[1] The reasoning is due to [36]

where $\mathcal{T}'_\alpha$ is a derivative of $\mathcal{T}_\alpha$ and $\omega$ is the step length.

For a comprehensive treatment and for references to the extensive literature on regularization of ill-posed problems, we refer the reader to the classical book [42], which has served here as a main source of inspiration. The appendix contains a few facts from the related variational calculus.

## 1.3 Overview and contributions

The core of the thesis consists of two parts:

- The first part is devoted to an analysis of parabolic boundary value problems with a certain class of nonstandard boundary conditions. Since nonstandard boundary conditions can be used in different situations, we do not restrict ourselves to one physical model. We establish unique solvability of the boundary value problems in appropriate spaces by monotone operator theory and Rothe's method. We make time and space discretization error analysis as well. We then compute the solutions numerically using the finite element method and perform some numerical experiments.

  The first part comprises Chapter 2, 3 and 4. In particular,

  - Chapter 2 discusses a boundary value problem with a nonlinear dynamical boundary condition. The motivation for this problem comes from the modeling of rainfall infiltration through soil. The dynamical boundary condition, which contains the time derivative, incorporates a possible saturation effect on the soil surface.

    We make here just time discretization error analysis. The results were presented at ACOMEN 2011, the Fifth International Conference on Advanced Computational Methods in Engineering, in Liege and subsequently published in Journal of Computational and Applied Mathematics [118].

  - Chapter 3 deals with a nonlinear heat diffusion equation with dynamical boundary condition of diffusive type. Despite the different physical situation, the mathematical model studied here can be seen as a generalization of the one from Chapter 2. The boundary condition contains besides other terms the Laplace-Beltrami operator, which acts in the directions tangential to the boundary and may allow for heat flow along the boundary.

    The results has been submitted to the journal Applied Mathematics and Computations.

  - Chapter 4 is drawn from the article [117] published in Journal of Mathematical Analysis and Applications. We consider the eddy current approximation of the Maxwell equations along with nonlinear impedance-like boundary

condition. The boundary condition here can be thought as the counterpart for the one in Chapter 2, but with the curl operator.

- The second part is based on a joint article with Valdemar Melicher. It has been submitted to Inverse Problems journal.

  In the article, we present a new approach to convexification of the Tikhonov regularization using a continuation method strategy. We embed the original minimization problem into a one-parameter family of minimization problems. Both the penalty term and the minimizer of the Tikhonov functional become dependent on a continuation parameter. In this way we can independently treat two main roles of the regularization term, which are stabilization of the ill-posed problem and introduction of the a priori knowledge. For zero continuation parameter we solve a relaxed regularization problem, which stabilizes the ill-posed problem in a weaker sense. The problem is recast to the original minimization by the continuation method and so the a priori knowledge is enforced. We apply this approach in the context of topology-to-shape geometry identification, where it allows to avoid the convergence of gradient-based methods to a local minima. We present illustrative results for magnetic induction tomography, which is an example of PDE constrained inverse problem.

  The second part is divided into the two chapters:

  - In Chapter 5 we propose and analyze the continuation approach in general. We provide more or less standard results on well-posedness, stability ad convergence of the approach. Under a strong condition of uniqueness, we provide a local correctness result of the continuation approach (Theorem 5.1).
  - Chapter 6 deals with piecewise-constant parameter identification problems, which have been our motivation to study continuation methods in the context of Tikhonov regularization. We briefly review the relevant state of the art in this field. Then, we introduce topology-to-shape continuation method (TSCM). In Section 6.2 we apply the TSCM to magnetic induction tomography (MIT), which has many applications, e.g. in biomedical imaging and non-destructive testing of materials. The section contains the implementation of the TSCM and a few numerical experiments for MIT.

# Part I

# On nonstandard boundary conditions

# Chapter 2

# Rainfall infiltration with saturation effect

In this chapter we consider a boundary value problem with boundary condition containing a time derivative. Our motivation comes from the modelling of rainfall infiltration through soil.

## 2.1 Physical model

We begin by a brief description of the physical model. A deeper discussion of groundwater modelling can be found in the book [14]. We refer the reader to the article [109] for the boundary condition in question.

The water flow through a soil is described by Darcy's law. It states that the seepage velocity of the water $\boldsymbol{q}$ is indirectly proportional to the product of the hydraulic conductivity $K$ and the gradient of hydraulic head, which itself is the the sum of the hydrostatic potential $u$ and the gravitational potential $z$

$$\boldsymbol{q} = -K\nabla(u + z).$$

The formula was determined experimentally by Darcy, but it can be also derived from Navier-Stokes equations by means of homogenization.

Coupled with the mass balance principle, Darcy's law leads to the groundwater flow equation

$$\partial_t \theta(u) - \nabla \cdot \boldsymbol{q}(u, \nabla u) = 0, \tag{2.1}$$

where $\theta$ is the moisture content.

We are particulary interested in the phenomena on the surface of the soil. There, the seepage velocity of the water should equal to the rainfall rate $q_0$. One can prescribe the

Neumann boundary condition

$$-\boldsymbol{q}\cdot\boldsymbol{n} = q_0,$$

which says that the rain totally infiltrates into the soil.

The ground surface can, in practice, become saturated. The water does not infiltrate into the soil, but accumulates on its surface. There appear water puddles and ponds. The boundary condition above fails to describe partial infiltration.

One way how to capture this phenomenon and correct the imperfection of the model lies in introducing the ponding rate $\partial_t u$. The Neumann boundary condition is then refined by adding a ponding term

$$H(u)\partial_t u - \boldsymbol{q}\cdot\boldsymbol{n} = q_0. \tag{2.2}$$

The unit step function $H$

$$H(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u \leq 0, \end{cases}$$

forces the ponding rate $\partial_t u$ to vanish if $u$ is zero on the ground surface. Let the function $\beta$ be an antiderivative of $H$, $\beta(u) = \max\{u, 0\}$. The ponding term can be then written in a more compact form

$$H(u)\partial_t u = \partial_t \int_0^u H(s)\,\mathrm{d}s = \partial_t \beta(u).$$

Note that all the information about what happens above the ground is embodied into boundary condition (2.2).

In what follows, we study a nonlinear degenerate diffusion parabolic equation of the form (2.1) which is accompanied with the dynamical boundary condition

$$\partial_t \beta(u) + \boldsymbol{q}(u, \nabla u)\cdot\boldsymbol{n} = g.$$

We suppose that the function $\beta$ is nonlinear as well. We state an exact mathematical formulation and demonstrate its unique solvability. The proof of existence and uniqueness are based on the Rothe method for time discretization and the monotone operator theory.

## 2.2 Mathematical formulation

We continue by stating a precise mathematical formulation. Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$ be a bounded domain with the Lipschitz continuous boundary $\Gamma$ and the unit normal outward vector $\boldsymbol{n}$. The boundary $\Gamma$ consists of the three non-overlapping parts $\Gamma_T, \Gamma_D$ and $\Gamma_N$. The boundary $\Gamma_T$ is the top boundary part, on which the dynamical boundary condition will be considered. The boundary part $\Gamma_D$ with the positive measure $meas(\Gamma_D) > 0$

represents the bottom layer of the soil and the homogeneous Dirichlet boundary condition is there prescribed. We set zero Neumann boundary condition on the side parts $\Gamma_N$. In a one dimensional case we just skip the part $\Gamma_N$. This corresponds to the simplest case of rainfall infiltration along a vertical line, when horizontal diffusion effects are neglected.

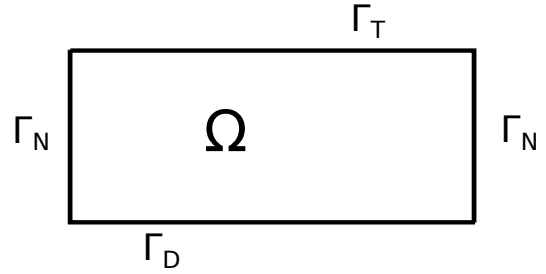



Figure 2.1: The domain $\Omega$

The complete boundary value problem for the unknown function $u = u(x,t)$ on the time interval $(0,T)$ reads as

$$\begin{aligned}
\partial_t \theta(u) - \nabla \cdot (\nabla u + \boldsymbol{b}(u)) &= f && \text{in } Q_T = \Omega \times (0,T) \\
\partial_t \beta(u) + (\nabla u + \boldsymbol{b}(u)) \cdot \boldsymbol{n} &= g && \text{on } \Gamma_T \times (0,T) \\
u &= 0 && \text{on } \Gamma_D \times (0,T) \\
(\nabla u + \boldsymbol{b}(u)) \cdot \boldsymbol{n} &= 0 && \text{on } \Gamma_N \times (0,T) \\
u(0) &= u_0 && \text{in } \Omega \text{ and on } \Gamma_T.
\end{aligned} \tag{2.3}$$

Settings with dynamical BCs have already been studied in the literature. A similar problem but with a linear differential operator was studied in [69]. The paper [70] deals with a nonlinear differential operator for $\beta = \theta$, however the case when $\beta'$ can vanish is not considered there. The nonlinearity there is, in fact, hidden under the differential operator. In the already mentioned article [109], the problem (2.3) was also discussed only with restriction to a Lipschitz-type nonlinearity. The boundary condition with a time derivative is sometimes called Wentzel boundary condition after A.D. Wentzel (see [121]). The authors of [48] have extensively studied similar boundary conditions using semigroup theory. We would like to mention the article [6] too, where the authors have studied a general setting for parabolic degenerate equations subjected to standard boundary conditions.

We now specify the assumptions under which we consider the problem (2.3). The nonlinear continuous functions $\theta$ and $\beta$ are monotonically increasing and their deriva-

tives can vanish or they can be unbounded

$$\begin{aligned} 0 \le \beta', \quad |\beta(s)| \le C(1+|s|), \\ 0 < \lambda \le \theta', \quad |\theta(s)| \le C(1+|s|). \end{aligned} \tag{2.4}$$

For convenience, we assume that

$$\beta(0) = 0 \quad \text{and} \quad \theta(0) = 0.$$

The function $\theta$ is moreover strictly monotone. Neither $\theta$ nor $\beta$ have any upper bound for its first derivative, hence they can degenerate in this sense. The continuous vector-valued function $\boldsymbol{b} : \mathbb{R} \to \mathbb{R}^d$ is bounded together with its first derivative

$$|\boldsymbol{b}(u) = (b_1(u), \dots, b_d(u))| \le C, \quad |\boldsymbol{b'}(u)| = |(b_1'(u), \dots, b_d'(u))| \le C. \tag{2.5}$$

The input data $g$ and $f$ are functions of the space and the time variable. For the reasons which will become clear later we will consider them as the continuous functions from the time interval $[0, T]$ to the Lebesgue spaces $L^2(\Omega)$ and $L^2(\Gamma_T)$ respectively, i.e.

$$f \in C([0,T], L^2(\Omega)), \quad g \in C([0,T], L^2(\Gamma_T)). \tag{2.6}$$

Let us introduce some notation. We denote by $(u, v)$ the standard $L^2$-scalar product of the functions $u$ and $v$ on the domain $\Omega$ and by $\|u\|$ the corresponding norm, that is

$$(u,v) = \int_\Omega u(x)v(x)\,\mathrm{d}x, \qquad \|u\| = \sqrt{(u,u)}.$$

The integration over a part of boundary will be indicated by a relevant subscript, e.g.

$$(u,v)_{\Gamma_T} = \int_{\Gamma_T} u(s)v(s)\,\mathrm{d}s, \qquad \|u\|_{\Gamma_T} = \sqrt{(u,u)_{\Gamma_T}}.$$

The symbol $H^1(\Omega)$ will stand for the standard Sobolev space with square integrable first derivatives (see [77] for more details)

$$H^1(\Omega) = \{u \in L^2(\Omega) : \nabla u \in [L^2(\Omega)]^d\},$$

which is equipped with the norm $\|u\|_{H^1(\Omega)}$

$$\|u\|_{H^1(\Omega)} = \sqrt{\|u\|^2 + \|\nabla u\|^2}.$$

We note already here that throughout the thesis, as it is usual in the analysis of this sort, the letters $C, \varepsilon$ and $C_\varepsilon$ will denote generic positive constants. They depend only on a priori known quantities, where $\varepsilon$ is sufficiently small and $C_\varepsilon$ is large.

To derive a weak formulation of (2.3) we multiply the first equation by a test function $\varphi = \varphi(x)$ from a suitable function space $V$. Integrating by parts over the domain $\Omega$ and using the boundary conditions produces the weak formulation which is required to hold true for almost every time $t \in (0,T)$. A weak solution belongs to the space $L^2((0,T),V)$ consisting of all measurable functions $u : [0,T] \to V$ with the norm

$$\|u\|_{L^2([0,T],V)} = \sqrt{\int_0^T \|u\|_V^2 \, \mathrm{d}t}.$$

A natural choice of the test space $V$ in our case is the following one

$$V = \{\varphi \in H^1(\Omega) : \quad \varphi|_{\Gamma_D} = 0\} \quad \text{with the norm } \|\varphi\|_V = \|\varphi\|_{H^1(\Omega)}. \tag{2.7}$$

Its dual space is denoted by $V^*$. It is worth noting that all the functions $\varphi \in V$ satisfy the Friedrichs inequality

$$C \|\varphi\|_{H^1(\Omega)} \leq \|\nabla\varphi\|. \tag{2.8}$$

The above inequality establishes the equivalence between Sobolev and gradient norm on the space $V$. We will frequently use this property.

The weak formulation of (2.3) consequently reads as:
find $u \in L^2((0,T),V)$ and $\partial_t\theta(u), \partial_t\beta(u) \in L^2((0,T),V^*)$ such that the equation

$$(\partial_t\theta(u),\varphi) + (\nabla u + \boldsymbol{b}(u), \nabla\varphi) + (\partial_t\beta(u),\varphi)_{\Gamma_T} = (f,\varphi) + (g,\varphi)_{\Gamma_T} \tag{2.9}$$

holds true for all $\varphi \in V$ and a.e in $(0,T)$.

The function $u$ is called a (weak) solution of (2.3). In this chapter we establish the unique solvability of the problem (2.9).

## 2.3 Stability

This section deals with stability results for the time discretization of the problem (2.9). They will be needed in the next section when proving existence and uniqueness of a solution to this problem.

We discretize the continuous problem (2.9) in time. The time interval $[0,T]$ is divided into $n \in \mathbb{N}$ equidistant subintervals $[t_{i-1}, t_i]$ for $t_i = i\tau$, where $\tau = \frac{T}{n}$. Adopting the standard notation for a discretized function and its backward difference,

$$w_i = w(t_i) \quad \text{and} \quad \delta w_i = \frac{w_i - w_{i-1}}{\tau}, \tag{2.10}$$

we approximate the original problem by the sequence of steady BVPs

$$\begin{aligned}(\delta\theta(u_i),\varphi) + (\nabla u_i + \boldsymbol{b}(u_i), \nabla\varphi) + (\delta\beta(u_i),\varphi)_{\Gamma_T} \\ = (f_i,\varphi) + (g_i,\varphi)_{\Gamma_T} \qquad \forall\varphi \in V\end{aligned} \tag{2.11}$$

for $i = 1, \ldots, n$.

The first lemma asserts the existence and uniqueness of the solution to (2.11) on every time step.

**Lemma 2.1.** *Suppose (2.4)-(2.6), $u_0 \in V$. Then there exist $\tau_0 > 0$ and a unique $u_i \in V$ solving the variational problem (2.11) for any $i = 1, \ldots, n$ and $\tau < \tau_0$.*

*Proof.* We apply the theory of monotone operators. The reader can consult appendix for the main theorem on monotone operators, which we will use.

Let us first define the mapping $A$ from the space $V$ to its dual $V^*$

$$A : V \to V^*, \qquad u \mapsto A(u).$$

The action of the image $A(u)$ on a function $\varphi \in V$ is defined by the left-hand side (LHS) of the formula (2.11) minus the terms from the previous time step, which are supposed to be known

$$\langle A(u), \varphi \rangle = (\theta(u), \varphi) + \tau \left(\nabla u + \boldsymbol{b}(u), \nabla \varphi\right) + \left(\beta(u), \varphi\right)_{\Gamma_T}, \qquad \varphi \in V.$$

The mapping $A$ is hemicontinuous. This can be verified by using the continuity of the functions $\beta, \theta$ and $\boldsymbol{b}$.

The monotonicity of $\beta$ and $\theta$ ensures the strict monotonicity of $A$, i.e.

$$\langle A(u) - A(v), u - v \rangle \geq C \|u - v\|^2_{H^1(\Omega)}.$$

Indeed, it follows from mean value theorem that

$$\begin{aligned}
\langle A(u) - A(v), u - v \rangle =& (\theta(u) - \theta(v), u - v) + \tau \left\|\nabla(u - v)\right\|^2 \\
&+\tau \left(\boldsymbol{b}(u) - \boldsymbol{b}(v), \nabla(u - v)\right) + \left(\beta(u) - \beta(v), u - v\right)_{\Gamma_T} \\
=& (\theta'(\xi_1)[u - v], u - v) + \tau \|\nabla(u - v)\|^2 \\
&+\tau(\boldsymbol{b}'(\xi_2)[u - v], \nabla(u - v)) + \left(\beta'(\xi_3)[u - v], u - v\right)_{\Gamma_T} \\
\geq& \lambda \|u - v\|^2 + \tau \|\nabla(u - v)\|^2 - \tau C \|u - v\| \|\nabla(u - v)\|.
\end{aligned}$$

If the time step $\tau$ is sufficiently small, then

$$\begin{aligned}
\langle A(u) - A(v), u - v \rangle \geq& (\lambda - \tau C)\|u - v\|^2 + \frac{\tau}{2}\|\nabla(u - v)\|^2 \\
\geq& \frac{\tau}{2}\|\nabla(u - v)\|^2 \geq \frac{\tau}{2}\|u - v\|^2_{H^1(\Omega)},
\end{aligned}$$

where the Friedrichs inequality (2.8) has been invoked.

The mapping $A$ is coercive, i.e.

$$\frac{\langle A(u), u \rangle}{\|u\|_{H^1(\Omega)}} \to \infty \quad \text{for } \|u\|_{H^1(\Omega)} \to \infty,$$

This can be seen from the following lower bound

$$\begin{aligned}\langle A(u), u\rangle &= (\theta(u), u) + \tau\left(\nabla u + \boldsymbol{b}(u), \nabla u\right) + (\beta(u), u)_{\Gamma_T} \\ &\geq \lambda\|u\|^2 + \tau\|\nabla u\|^2 - |\left(\boldsymbol{b}(u), \nabla u\right)| \\ &\geq \lambda\|u\|^2 + \tau\|\nabla u\|^2 - C_\varepsilon - \varepsilon\|\nabla u\|^2 \\ &\geq C_1\|u\|^2_{H^1(\Omega)} - C_2,\end{aligned}$$

where the $\varepsilon$-Young inequality has been applied.

In light of Theorem A.10, we conclude that there exists the unique $u_i$ for any $i = 1, \ldots, n$ solving

$$\langle A(u_i), \varphi\rangle = (\theta(u_{i-1}), \varphi) + (\beta(u_{i-1}), \varphi)_{\Gamma_T} + \tau\big[\,(f_i, \varphi) + (g_i, \varphi)_{\Gamma_T}\,\big]$$

for any $\varphi \in V$, which had to be demonstrated. □

Let us note that the problem (2.11) is nonlinear at each time step. For suitable linearizations we refer the reader to e.g. [72, 101].

The next two lemmas are the so-called a priori estimates for $u_i$ and $\delta u_i$. They provide a bound for the size of solution and its derivatives. This bound is in fact independent of the time discretization, in other words it is independent of $n \in \mathbb{N}$.

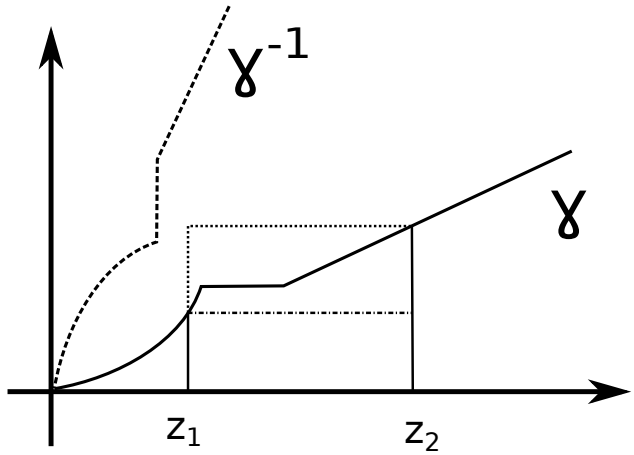



Figure 2.2: Proof of (2.12) without words for $\gamma$ and its inverse "graph" $\gamma^{-1}$

We first make a simple observation, which will turn out very useful in the proofs below. Let $\gamma$ be a monotone increasing real function with $\gamma(0) = 0$. Define the potential

of $g$ as

$$\Phi_\gamma(z) := \int_0^z \gamma(s)\,\mathrm{d}s.$$

Then, the inequality

$$\gamma(z_1)(z_2 - z_1) \le \Phi_\gamma(z_2) - \Phi_\gamma(z_1) \le \gamma(z_2)(z_2 - z_1) \tag{2.12}$$

holds for any $z_1, z_2 \in \mathbb{R}$. Instead of a formal proof we offer a graphic demonstration on Figure 2.2. The particular terms in the inequality represent the signed areas under the dot-dashed, full and dotted line respectively. The assertion (2.12) then becomes self-evident. One can generalize this assertion to monotone increasing graphs like in the same figure, cf. [102].

**Lemma 2.2.** *Suppose (2.4)-(2.6), $u_0 \in V$. If $u_i$ is the solution of (2.11), then there exists a positive constant C such that*

$$\sum_{i=1}^{n} \|\nabla u_i\|^2 \tau \le C.$$

*Proof.* $(i)$ The proof is straightforward. We set $\varphi = \tau u_i$ in (2.11) and add it up for $i = 1, \ldots, n$ to obtain

$$\begin{aligned} &\sum_{i=1}^{n} (\delta\theta(u_i), u_i)\,\tau + \sum_{i=1}^{n} \|\nabla u_i\|^2 \tau + \sum_{i=1}^{n} (\boldsymbol{b}(u_i), \nabla u_i)\,\tau \\ &+ \sum_{i=1}^{n} (\delta\beta(u_i), u_i)_{\Gamma_T}\,\tau = \sum_{i=1}^{n} (f_i, u_i)\,\tau + \sum_{i=1}^{n} (g_i, u_i)_{\Gamma_T}\,\tau. \end{aligned} \tag{2.13}$$

We bound term by term the LHS of (2.13) from below. Let $\tilde{\Phi}_\theta(u) = u\theta(u) - \Phi_\theta(u)$, which is positive and monotonically increasing. According to (2.12), we have for the

first sum

$$\begin{aligned}
&\sum_{i=1}^{n}(\theta(u_i)-\theta(u_{i-1}),u_i)\\
&=(\theta(u_n),u_n)-(\theta(u_0),u_0)-\sum_{i=1}^{n}(u_i-u_{i-1},\theta(u_{i-1}))\\
&\geq(\theta(u_n),u_n)-(\theta(u_0),u_0)-\sum_{i=1}^{n}\int_\Omega[\Phi_\theta(u_i)-\Phi_\theta(u_{i-1})]\\
&=\left[(\theta(u_n),u_n)-\int_\Omega\Phi_\theta(u_n)\right]-\left[(\theta(u_0),u_0)-\int_\Omega\Phi_\theta(u_0)\right]\\
&=\int_\Omega\left[\tilde{\Phi}_\theta(u_n)-\tilde{\Phi}_\theta(u_0)\right]\\
&\geq-\int_\Omega u_0\theta(u_0)\\
&\geq -C.
\end{aligned}$$

The last inequality is legitimate because of the growth condition for $\theta$. The sum with $\beta$ can be estimated in the same way

$$\sum_{i=1}^{n}(\delta\beta(u_i),u_i)_{\Gamma_T}\,\tau\geq -C.$$

A lower bound for the advection sum in (2.13) is found by the Cauchy and $\varepsilon$-Young inequality

$$\begin{aligned}
\sum_{i=1}^{n}(\boldsymbol{b}(u_i),\nabla u_i)\,\tau&\geq-\sum_{i=1}^{n}|\,(\boldsymbol{b}(u_i),\nabla u_i)\,|\tau\geq\sum_{i=1}^{n}-\|\boldsymbol{b}(u_i)\|\,\|\nabla u_i\|\,\tau\\
&\geq\sum_{i=1}^{n}-\big(\frac{1}{4\varepsilon}\,\|\boldsymbol{b}(u_i)\|^2+\varepsilon\,\|\nabla u_i\|^2\big)\tau\geq-C_\varepsilon-\varepsilon\sum_{i=1}^{n}\|\nabla u_i\|^2\,\tau.
\end{aligned}$$

We now estimate the right hand side (RHS) of (2.13) from above. Aconsecutive application of the Cauchy, Young and Friedrichs inequalities and the trace theorem shows that

$$\sum_{i=1}^{n}(f_i,u_i)\,\tau+\sum_{i=1}^{n}(g_i,u_i)_{\Gamma_T}\,\tau\leq C_\varepsilon+\varepsilon\sum_{i=1}^{n}\|\nabla u_i\|^2\,\tau.$$

We write down all the partial results to see that

$$(1-2\varepsilon)\sum_{i=1}^{n}\|\nabla u_i\|^2\,\tau\leq C_\varepsilon.$$

The proof is complete by choosing $\varepsilon$ small enough. □

**Lemma 2.3.** *Suppose (2.4)-(2.6), $u_0 \in V$. If $u_i$ is the solution of (2.11), then there exists a positive constant C such that*

$$\sum_{i=1}^{n} \|\delta u_i\|^2 \tau + \max_{1\le j\le n} \|\nabla u_j\|^2 + \sum_{i=1}^{n} \|\nabla(u_i - u_{i-1})\|^2 \le C.$$

*Proof.* Proceeding similarly to the previous proof, we set $\varphi = \tau \delta u_i$ and sum it over $i = 1, \dots, j$ to get

$$\begin{aligned} &\sum_{i=1}^{j} (\delta\theta(u_i), \delta u_i)\, \tau + \sum_{i=1}^{j} (\nabla u_i, \nabla \delta u_i)\, \tau + \sum_{i=1}^{j} (\boldsymbol{b}(u_i), \nabla \delta u_i)\, \tau \\ &+ \sum_{i=1}^{j} (\delta\beta(u_i), \delta u_i)_{\Gamma_T}\, \tau = \sum_{i=1}^{j} (f_i, \delta u_i)\, \tau + \sum_{i=1}^{j} (g_i, \delta u_i)_{\Gamma_T}\, \tau. \end{aligned} \tag{2.14}$$

The Abel summation rule for the gradient sum reveals that

$$\sum_{i=1}^{j} 2\,(\nabla u_i, \nabla \delta u_i)\, \tau = \|\nabla u_j\|^2 + \sum_{i=1}^{j} \|\nabla u_i - \nabla u_{i-1}\|^2 - \|\nabla u_0\|^2 .$$

The advection sum on the LHS of (2.14) can be rewritten by the same summation rule

$$\sum_{i=1}^{j} (\boldsymbol{b}(u_i), \delta \nabla u_i)\, \tau = (\boldsymbol{b}(u_j), \nabla u_j) - (\boldsymbol{b}(u_0), \nabla u_0) - \sum_{i=1}^{j} (\delta \boldsymbol{b}(u_i), \nabla u_{i-1})\, \tau.$$

We see in light of $\varepsilon$-Young's inequality and the assumption (2.5) that

$$\sum_{i=1}^{j} (\boldsymbol{b}(u_i), \delta \nabla u_i)\, \tau \le \varepsilon \|\nabla u_j\|^2 + C_\varepsilon + \varepsilon \sum_{i=1}^{j} \|\delta u_i\|^2 \tau.$$

The monotonicity argument for $\theta$ and $\beta$ yields

$$\sum_{i=1}^{j} (\delta\theta(u_i), \delta u_i)\, \tau + \sum_{i=1}^{j} (\delta\beta(u_i), \delta u_i)_{\Gamma_T}\, \tau \ge \lambda \sum_{i=1}^{j} \|\delta u_i\|^2 \tau.$$

For the terms on the RHS of (2.14) we consecutively apply the Abel summation; Cauchy's, Young's inequalities; the trace theorem and the Friedrichs inequality. Let us display it

just for the sum containing the function $g$

$$\begin{aligned}\sum_{i=1}^{j}(g_i,\delta u_i)_{\Gamma_T}\tau &= (g_j,u_j)_{\Gamma_T}-(g_0,u_0)_{\Gamma_T}-\sum_{i=1}^{j}(\delta g_i,u_{i-1})_{\Gamma_T}\tau\\ &\le \|g_j\|_{\Gamma_T}\|u_j\|_{\Gamma_T}+\|g_0\|_{\Gamma_T}\|u_0\|_{\Gamma_T}+\sum_{i=1}^{j}\|\delta g_i\|_{\Gamma_T}\|u_{i-1}\|_{\Gamma_T}\tau\\ &\le C_\varepsilon+\varepsilon\|u_j\|_{\Gamma}+C+\sum_{i=1}^{j}\Big(\|\delta g_i\|^2_{\Gamma_T}+\|u_{i-1}\|^2_{\Gamma_T}\Big)\tau\\ &\le \varepsilon\|\nabla u_j\|^2+C_\varepsilon.\end{aligned}$$

The sum with $u_{i-1}$ has been bounded by the a priori estimate from Lemma 2.2. Collecting all the estimates we arrive at

$$\begin{aligned}\lambda\sum_{i=1}^{j}\|\delta u_i\|^2\tau+\frac{1}{2}\left[\|\nabla u_j\|^2+\sum_{i=1}^{j}\|\nabla u_i-\nabla u_{i-1}\|^2-\|\nabla u_0\|^2\right]\\ \le \varepsilon\|\nabla u_j\|^2+C_\varepsilon+\varepsilon\sum_{i=1}^{j}\|\delta u_i\|^2\tau.\end{aligned}$$

We pick a sufficiently small positive $\varepsilon<\min\{1/2,\lambda\}$ and take the maximum over the index $j$ to conclude the proof. □

## 2.4 The Rothe method

Having proved the unique solvability of (2.11) on every time step, we direct our attention back to the problem on the whole interval. We use the Rothe method to demonstrate existence of a unique solution of the problem (2.9) . Let $u_n$ be the following piecewise-linear-in-time function

$$\begin{aligned}&u_n(0)=u(0),\\ &u_n(t)=u_{i-1}+\delta u_i(t-t_{i-1})\quad \text{for}\, t\in(t_{i-1},t_i],\end{aligned}$$

and $\overline{u}_n$ be the constant-in-time function

$$\begin{aligned}&\overline{u}_n(0)=u(0),\\ &\overline{u}_n(t)=u_i\quad \text{for}\, t\in(t_{i-1},t_i].\end{aligned}$$

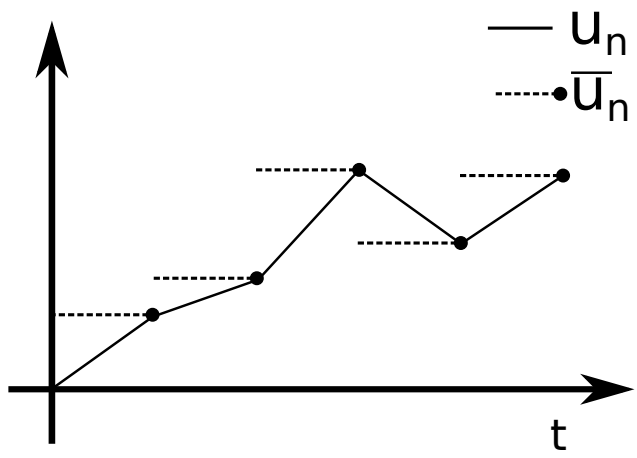



Figure 2.3: The Rothe functions

The piecewise-linear-in-time counterparts $\theta_n$ and $\beta_n$ of the functions $\theta(u)$ and $\beta(u)$ respectively are defined in accordance with it

$$\begin{aligned} &\theta_n(0) = \theta(u(0)), \qquad \theta_n(t) = \theta(u_{i-1}) + \delta\theta(u_i)(t - t_{i-1}) \qquad \text{for } t \in (t_{i-1}, t_i], \\ &\beta_n(0) = \beta(u(0)), \qquad \beta_n(t) = \beta(u_{i-1}) + \delta\beta(u_i)(t - t_{i-1}) \qquad \text{for } t \in (t_{i-1}, t_i]\,. \end{aligned}$$

The data functions $f$ and $g$ are also discretized

$$\overline{f}_n(t) = f_i \quad \text{and} \quad \overline{g}_n(t) = g_i \qquad \text{for } t \in (t_{i-1}, t_i].$$

The new notation enables us to rewrite the sequence (2.11) for $i = 1, \ldots, n$ in the more compact form

$$\begin{aligned} (\partial_t \theta_n, \varphi) + (\nabla \overline{u}_n + \boldsymbol{b}(\overline{u}_n), \nabla\varphi) + (\partial_t \beta_n, \varphi)_{\Gamma_T} \\ = \left(\overline{f}_n, \varphi\right) + (\overline{g}_n, \varphi)_{\Gamma_T} \qquad \forall \varphi \in V. \end{aligned} \tag{2.15}$$

This notation suggests that taking $n \to \infty$ suffices to recover the original problem. The first theorem proves that the sequences $u_n$ and $\overline{u}_n$ do converge in some sense to a function $u$ which is the unique solution of (2.9).

**Theorem 2.1** (Existence)**.** *Suppose (2.4)-(2.6), $u_0 \in V$. Then there exists a solution to (2.9).*

*Proof.* We begin by stating all the partial results which we will use when passing to the limit $n \to \infty$.

Firstly, it follows easily from Lemma 2.2 and the Friedrichs inequality that the Rothe sequence $\{\overline{u}_n\}_{n\in\mathbb{N}}$ is bounded in the space $L^2((0,T),V)$

$$\begin{aligned}\|\overline{u}_n\|_{L^2((0,T),V)} &= \sqrt{\int_0^T \|\overline{u}_n\|^2_{H^1(\Omega)}\,\mathrm{d}t} = \sqrt{\sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|\overline{u}_n\|^2_{H^1(\Omega)}\,\mathrm{d}t}\\ &= \sqrt{\sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|u_i\|^2_{H^1(\Omega)}\,\mathrm{d}t} = \sqrt{\sum_{i=1}^n \|u_i\|^2_{H^1(\Omega)}\,\tau}\\ &\le \sqrt{\sum_{i=1}^n C\,\|\nabla u_i\|^2\,\tau}\\ &\le C.\end{aligned}$$

Every bounded sequence in a reflexive space contains a weakly convergent subsequence. The space $L^2((0,T),V)$ is reflexive and so

$$\overline{u}_n \rightharpoonup u \qquad \text{in } L^2((0,T),V). \tag{2.16}$$

For simplicity of notation all the subsequences are denoted by the same subscript as the initial sequence. Lemma 2.3 furthermore shows that the sequences $\overline{u}_n$ and $u_n$ share the common limit $u$ in the same space $L^2((0,T),V)$

$$\begin{aligned}\lim_{n\to\infty} \|u_n-\overline{u}_n\|^2_{L^2((0,T),V)} &= \lim_{n\to\infty} \int_0^T \|\overline{u}_n-u_n\|^2_{H^1(\Omega)}\,\mathrm{d}t\\ &= \lim_{n\to\infty} \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|u_i-u_{i-1}-(t-t_{i-1})\delta u_i\|^2_{H^1(\Omega)}\,\mathrm{d}t\\ &\le \lim_{n\to\infty} \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|2(u_i-u_{i-1})\|^2_{H^1(\Omega)}\,\mathrm{d}t\\ &\le \lim_{n\to\infty} C\sum_{i=1}^n \|\nabla(u_i-u_{i-1})\|^2\,\tau \le \lim_{n\to\infty} \frac{C}{n}\\ &=0.\end{aligned}$$

We continue by showing that the sequence $\{\overline{u}_n\}_{n\in\mathbb{N}}$ is 2-mean equicontinuous in $L^2(Q_T)$. That is to say that for every $\varepsilon>0$ there exists a $\delta>0$ such that

$$\int_0^T \int_\Omega |\overline{u}_n(x+h,t+s)-\overline{u}_n(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t < \varepsilon^2 \tag{2.17}$$

for each $n \in \mathbb{N}$ and $t \in \mathbb{R}, h \in \mathbb{R}^d$ with $\sqrt{|t|^2 + |h|^2} < \delta$. To show it we split the integrand into two parts by the triangle inequality

$$\begin{aligned} &|\overline{u}_n(x+h,t+s) - \overline{u}_n(x,t)| \\ \leq\ &|\overline{u}_n(x+h,t+s) - \overline{u}_n(x,t+s)| + |\overline{u}_n(x,t+s) - \overline{u}_n(x,t)|. \end{aligned}$$

Suppose that the differences are $t$ and $h$ are sufficiently small so that the function $\overline{u}_n$ can be extended outside the $\Omega \times (0,T)$. For the space difference we get by the Hadamard lemma (see Lemma A.7) and Lemma 2.2

$$\begin{aligned} &\int_0^T \int_\Omega |\overline{u}_n(x+h,t+s) - \overline{u}_n(x,t+s)|^2 \,\mathrm{d}x\,\mathrm{d}t \\ =&\int_0^T \int_\Omega \Big| \int_0^1 \nabla \overline{u}_n(x+\vartheta h, t+s) h \,\mathrm{d}\vartheta \Big|^2 \,\mathrm{d}x\,\mathrm{d}t \\ \leq&\int_0^T \int_\Omega \int_0^1 |\nabla \overline{u}_n(x+\vartheta h, t+s) h|^2 \,\mathrm{d}\vartheta\,\mathrm{d}x\,\mathrm{d}t \\ \leq&\int_0^T \int_\Omega |\nabla \overline{u}_n(x,t+s)|^2 |h|^2 \,\mathrm{d}x\,\mathrm{d}t \\ \leq&\sum_{i=1}^n \|\nabla u_i\|^2 |h|^2 \tau \leq C|h|^2. \end{aligned}$$

The handling of the time difference is more complicated. We will establish, instead of the definition (2.17), the mean equicontinuity in time in terms of sequences. Introducing $k \in \mathbb{N}$ with $s = \frac{T}{k}$, we will verify that

$$\lim_{k\to\infty} \int_0^T \|\overline{u}_n(t+s) - \overline{u}_n(t)\|^2 \,\mathrm{d}t = 0 \qquad \text{uniformly for every function } \overline{u}_n.$$

We assume that that $\tau m = s$ for $m \in \mathbb{N}$ (i.e. $n = mk$ and $\tau = \frac{T}{mk}$) to bypass some technicalities. It can be deduced in this case that

$$\begin{aligned} \int_0^T \|\overline{u}_n(t+s) - \overline{u}_n(t)\|^2 \,\mathrm{d}t &= \sum_{i=1}^{mk} \int_{t_{i-1}}^{t_i} \|\overline{u}_n(t+s) - \overline{u}_n(t)\|^2 \,\mathrm{d}t \\ &= \sum_{i=1}^{mk} \|u_{m+i} - u_i\|^2 \frac{T}{mk} \leq \sum_{j=1}^{\infty} \|u_{j+1} - u_j\|^2 \frac{T}{k} \\ &\leq \sum_{j=1}^{\infty} C \|\nabla(u_{j+1} - u_j)\|^2 \frac{T}{k} \\ &\leq \frac{C}{k} \to 0 \quad \text{ as } k \to \infty. \end{aligned}$$

The trick here has been to choose another time discretization such that $u_{i+m} = u_{j+1}$ and utilize the fact that the constant in Lemma 2.2 does not depend on any time discretization

$$\sum_{j=1}^{k} \|\nabla(u_{j+1} - u_j)\|^2 \frac{T}{k} \leq C.$$

Since the sequence $\overline{u}_n$ is bounded and 2-mean equicontinuous, we conclude from the Riesz-Kolmogorov theorem (see Theorem A.19) that it is also relatively compact in $L^2(Q_T)$. There exists therefore a subsequence such that

$$\lim_{n\to\infty} \overline{u}_n \to u \qquad \text{in } L^2(Q_T).$$

The convergence in norm implies the convergence almost everywhere for a subsequence

$$\overline{u}_n \to u \qquad \text{a.e. in } Q_T.$$

Further, the inequality

$$\|w\|_{\Gamma}^2 \leq \varepsilon \|\nabla w\|^2 + C_\varepsilon \|w\|^2$$

holds true for any function $w \in H^1(\Omega)$ and $0 < \varepsilon < \varepsilon_0$ according to the book [86]. Let us insert $w = \overline{u}_n - u$ and employ (2.16) to get

$$\begin{aligned}\int_0^T \|\overline{u}_n - u\|_{\Gamma_T}^2 &\leq \varepsilon \int_0^T \|\nabla(\overline{u}_n - u)\|^2 + C_\varepsilon \int_0^T \|\overline{u}_n - u\|^2 \\ &\leq C\varepsilon + C_\varepsilon \int_0^T \|\overline{u}_n - u\|^2 .\end{aligned}$$

Taking the limit for $\tau \to 0$ leads to

$$\lim_{\tau\to 0} \int_0^T \|\overline{u}_n - u\|_{\Gamma_T}^2 \leq C\varepsilon \qquad \text{for any small } \varepsilon > 0$$

which is nothing more then the convergence in norm. For this reason the subsequence $\overline{u}_n$ converges almost everywhere on the boundary part $\Gamma_T$ too,

$$\overline{u}_n \to u \quad \text{a.e. in } \Gamma_T \times (0,T).$$

The time derivative $\partial_t u_n$ is also bounded

$$\begin{aligned}\|\partial_t u_n\|_{L^2((0,T),L^2(\Omega))} &= \int_0^T \|\partial_t u_n\|^2 \,\mathrm{d}t \\ &= \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \|\partial_t u_n\|^2 \,\mathrm{d}t = \sum_{i=1}^{n} \|\delta u_i\|^2 \tau \\ &\leq C,\end{aligned}$$

and hence $\partial_t u_n \rightharpoonup z \in L^2((0,T),L^2(\Omega))$. It holds on the other hand for any $\varphi \in V$ that

$$(u_n(t),\varphi) - (u_0,\varphi) = \int_0^t (\partial_s u_n,\varphi)\,\mathrm{d}s.$$

Passing to the limit $n\to\infty$ for arbitrary $t\in[0,T]$ in the above identity yields

$$(u(t),\varphi) - (u_0,\varphi) = \int_0^t (z,\varphi)\,\mathrm{d}s,$$

which means that $z=\partial_t u$.

We claim next that the interchange of limits

$$\begin{aligned}\lim_{n\to\infty} (\theta_n(t),\varphi) + (\beta_n(t),\varphi)_{\Gamma_T} &= \lim_{n\to\infty} (\theta(\overline{u}_n(t)),\varphi) + (\beta(\overline{u}_n(t)),\varphi)_{\Gamma_T}\\ &= (\theta(u(t)),\varphi) + (\beta(u(t)),\varphi)_{\Gamma_T}\end{aligned}$$

holds true. Indeed, rewriting the Rothe formulation (2.15) for given $t\in(t_{i-1},t_i]$ reveals that

$$\begin{aligned}&(\theta_n(t)-\theta(\overline{u}_n(t)),\varphi) + (\beta_n(t)-\beta(\overline{u}_n(t)),\varphi)_{\Gamma_T}\\ =&\frac{t-t_i}{\tau}\left[(\theta(\overline{u}_n(t))-\theta(\overline{u}_n(t-\tau)),\varphi) + (\beta(\overline{u}_n(t))-\beta(\overline{u}_n(t-\tau)),\varphi)_{\Gamma_T}\right]\\ =&(t-t_i)\left[\left(\overline{f}_n(t),\varphi\right) + (\overline{g}_n(t),\varphi)_{\Gamma_T} - (\nabla\overline{u}_n(t)+\boldsymbol{b}(\overline{u}_n(t)),\nabla\varphi)\right].\end{aligned}$$

The assumptions (2.5), (2.6) and the a priori estimates now ensure that the right side diminishes as $n$ increases

$$\begin{aligned}&\left|\,(\theta_n(t)-\theta(\overline{u}_n(t)),\varphi) + (\beta_n(t)-\beta(\overline{u}_n(t)),\varphi)_{\Gamma_T}\,\right|\\ &\qquad\le \tau C\|\varphi\|_{H^1(\Omega)} \to 0 \quad \text{for } n\to\infty,\end{aligned} \tag{2.18}$$

which is desired result.

Finally, in view of the above considerations, we may apply the Lebesgue dominated theorem A.15 to pass to the limit for $n\to\infty$ in (2.15). We arrive at

$$\begin{aligned}&(\theta(u(t)),\varphi) + (\beta(u(t)),\varphi)_{\Gamma_T} - (\theta(u_0),\varphi) + (\beta(u_0),\varphi)_{\Gamma_T}\\ =&\int_0^t \left[(f(s),\varphi) + (g(s),\varphi)_{\Gamma_T} - (\nabla u(s)+\boldsymbol{b}(u(s)),\nabla\varphi)\right]\,\mathrm{d}s.\end{aligned}$$

The above equality is valid for any $t\in(0,T)$ and so differentiation with respect to the time variable proves that $u$ is a solution to (2.9).

Let us note here that the term $\partial_t u$ on the boundary part $\Gamma_T$ is well defined despite the fact that we have not provided any estimate for it. This term is in fact implicitly defined by the above variational formulation. Therefore, it can be bounded by using this formulation. □

The solution $u$ is moreover unique as the theorem below demonstrates.

**Theorem 2.2** (Uniqueness)**.** *Suppose (2.4)-(2.6) and $u_0 \in V$. Then (2.9) admits at most one solution.*

*Proof.* We prove the theorem by contradiction. Assume that both $u$ and $v$ solve (2.9). We subtract the corresponding variational formulations from each other, integrate in time, set the difference of both solutions as a test function and again integrate in time. This produces

$$\begin{aligned}\int_0^t (\theta(u)-\theta(v),u-v)+\int_0^t (\beta(u)-\beta(v),u-v)_{\Gamma_T}&\\ +\int_0^t\left(\int_0^s \nabla[u-v],\nabla[u(s)-v(s)]\right)&\\ =\int_0^t\left(\int_0^s \boldsymbol{b}(v)-\boldsymbol{b}(u),\nabla[u(s)-v(s)]\right).&\end{aligned} \tag{2.19}$$

The first line in (2.19) is non-negative by monotonicity argument

$$\int_0^t (\theta(u)-\theta(v),u-v)+\int_0^t (\beta(u)-\beta(v),u-v)_{\Gamma_T}\geq \lambda\int_0^t \|u-v\|^2.$$

Integration by parts yields

$$\begin{aligned}&\int_0^t\left(\int_0^s \nabla[u-v],\nabla[u(s)-v(s)]\right)\\ &=\left(\int_0^t \nabla[u-v],\int_0^t \nabla[u(s)-v(s)]\right)-\int_0^t\left(\nabla[u(s)-v(s)],\int_0^s \nabla[u-v]\right).\end{aligned}$$

The last term in (2.19) can be thus rewritten as

$$\int_0^t\left(\int_0^s \nabla[u-v],\nabla[u(s)-v(s)]\right)=\frac{1}{2}\left\|\int_0^t \nabla[u-v]\right\|^2.$$

We then observe that

$$\begin{aligned}\lambda\int_0^t \|u-v\|^2+\frac{1}{2}\left\|\int_0^t \nabla[u-v]\right\|^2 &\leq \left(\int_0^t \boldsymbol{b}(v)-\boldsymbol{b}(u),\int_0^t \nabla[u(s)-v(s)]\right)\\ &\quad-\int_0^t\left(\boldsymbol{b}(v)-\boldsymbol{b}(u),\int_0^s \nabla[u(s)-v(s)]\right)\\ &\leq \varepsilon\int_0^t \|u-v\|^2+C_\varepsilon\int_0^t\left\|\int_0^t \nabla[u-v]\right\|^2,\end{aligned}$$

where the integration by parts has been used for the advection term. We fix a sufficiently small positive $\varepsilon$ and by applying Gronwall's argument we arrive at

$$\left\| \int_0^t \nabla[u-v] \right\|^2 = 0.$$

The integral $\int_0^t [u-v]$ is constant as a function of the space variable for any time $t \in (0,T)$. Taking into account the fact that $u(t) = v(t)$ on $\Gamma_D$ we conclude that the two functions are identical, $u \equiv v$ in $\Omega$, which is the contradiction. □

**Remark 2.1.** *If the rainfall water is cumulated over the porous medium, then fully saturated zone inside the top layer of the porous medium can appear. The derivative $\theta'$ equals to zero and an elliptic equation takes place. The assumption (2.4) of Lemma 2 and 3 on strict monotonicity of $\theta$ is no longer satisfied. Even in this case one can establish a priori estimates for time derivatives using dual norms. Existence and uniqueness can be still showed, however no error estimates can be obtained.*

## 2.5 Error estimates

In this short section we address the question of the convergence rate of time discretization scheme. The importance of the following theorem lies in establishing the convergence bound which depends on the (decreasing) time step $\tau$.

**Theorem 2.3.** *Suppose (2.4)-(2.6), $u_0 \in V$. Then there exists a positive constant $C$ such that*

$$\int_0^T \|\overline{u}_n - u\|^2 + \left\| \int_0^T \nabla(\overline{u}_n - u) \right\|^2 \le C\tau\,.$$

*Proof.* We subtract the weak formulation (2.9) from (2.15). We set $\varphi(t) = -\int_s^t (\overline{u}_n(\sigma) - u(\sigma))\, d\sigma$ and integrate in time:

$$\begin{aligned} -\int_0^s \Big(\partial_t\theta_n - \partial_t\theta(u), \int_s^t (\overline{u}_n - u)\Big) - \int_0^s \Big(\nabla(\overline{u}_n - u), \int_s^t \nabla(\overline{u}_n - u)\Big) \\ - \int_0^s \Big(\partial_t\beta_n - \partial_t\beta(u), \int_s^t (\overline{u}_n - u)\Big)_{\Gamma_T} \\ = \int_0^s \Big(\boldsymbol{b}(\overline{u}_n) - \boldsymbol{b}(u), \int_s^t \nabla(\overline{u}_n - u)\Big). \end{aligned} \tag{2.20}$$

We examine the formula (2.20) term by term. The integration by parts for the first

integral gives us

$$
\begin{aligned}
&-\int_0^s \Big(\partial_t(\theta_n-\theta(u)),\int_s^t(\overline{u}_n-u)\Big)\\
&=\int_0^s \big(\theta_n-\theta(u),\overline{u}_n-u\big)-\Big(\theta_n-\theta(u),\int_s^t(\overline{u}_n-u)\Big)\Big|_0^s\\
&=\int_0^s \big(\theta_n\pm\theta(\overline{u}_n)-\theta(u),\overline{u}_n-u\big)\\
&\geq\int_0^s \lambda\|\overline{u}_n-u\|^2-\int_0^s\big(\theta(\overline{u}_n)-\theta_n,\overline{u}_n-u\big).
\end{aligned}
$$

It easy to check for the second term in (2.20) that

$$
-\int_0^s \Big(\nabla(\overline{u}_n-u),\int_s^t\nabla(\overline{u}_n-u)\Big)=\frac{1}{2}\Big\|\int_0^s\nabla(\overline{u}_n-u)\Big\|^2.
$$

The third term in (2.20) is estimated in the similar way as the first one

$$
-\int_0^s \Big(\partial_t(\beta_n-\beta(u)),\int_s^t(\overline{u}_n-u)\Big)_{\Gamma_T}\geq-\int_0^s\big(\beta(\overline{u}_n)-\beta_n,\overline{u}_n-u\big)_{\Gamma_T}.
$$

We apply Young's inequality and mean value theorem to the RHS of (2.20) to get

$$
\begin{aligned}
&\int_0^s \Big(\boldsymbol{b}(\overline{u}_n)-\boldsymbol{b}(u),\int_s^t\nabla(u-\overline{u}_n)\Big)\\
&\leq\varepsilon\int_0^s\|\boldsymbol{b}(\overline{u}_n)-\boldsymbol{b}(u)\|^2+C_\varepsilon\int_0^s\Big\|\int_s^t\nabla(u-\overline{u}_n)\Big\|^2\\
&\leq\varepsilon\int_0^s\|\overline{u}_n-u\|^2+C_\varepsilon\int_0^s\Big\|\int_s^t\nabla(u-\overline{u}_n)\Big\|^2.
\end{aligned}
$$

Re-collecting all the estimates and choosing $\varepsilon$ sufficiently small yield

$$
\begin{aligned}
&\int_0^s\|\overline{u}_n-u\|^2+\Big\|\int_0^s\nabla(\overline{u}_n-u)\Big\|^2\\
&\leq C\int_0^s\Big\|\int_s^t\nabla(u-\overline{u}_n)\Big\|^2\\
&+C\int_0^s\big(\theta(\overline{u}_n)-\theta_n,\overline{u}_n-u\big)+C\int_0^s\big(\beta(\overline{u}_n)-\beta_n,\overline{u}_n-u\big)_{\Gamma_T}\\
&\leq C\int_0^s\Big\|\int_t^s\nabla(u-\overline{u}_n)\Big\|^2+C\tau.
\end{aligned}
$$

The result (2.18) and the statement (2.16) have been used above to obtain

$$\int_0^s \big(\theta(\overline{u}_n) - \theta_n, \overline{u}_n - u\big) + \int_0^s \big(\beta(\overline{u}_n) - \beta_n, \overline{u}_n - u\big)_{\Gamma_T} \le C\tau.$$

Finally, the Gronwall argument implies the statement of the theorem. □

## 2.6 Numerical experiments

We close this chapter with some numerical experiments. In the first two examples, we study an absolute error in the $L^2(Q_T)$ sense

$$E = \sqrt{\int_0^T \|u_{exact} - u_{numerical}\|^2 \, \mathrm{d}t}.$$

The third example concerns a qualitative behaviour of a solution with the dynamical BC. We make a comparision to the model with Neumann BC (see the first section). The experiments have been performed using FEniCS software [78].

Our computational scheme follows the theoretical analysis. We use backward Euler method to discretize the problem (2.9) in time. For the space discretization we use the finite element method with the first-order Lagrange finite elements and a sufficiently fine grid to make the space error lower order than the time error. We apply the fixed-point method for solving nonlinear algebraic systems. In the experiments the domain $\Omega$ is the unit square $(0,1)^2$ in $xy$-plane. Its triangulation consists of 5000 triangles. The dynamical BC is prescribed on the top of domain: $\Gamma_T = \{0 \le x \le 1\,, y = 1\}$.

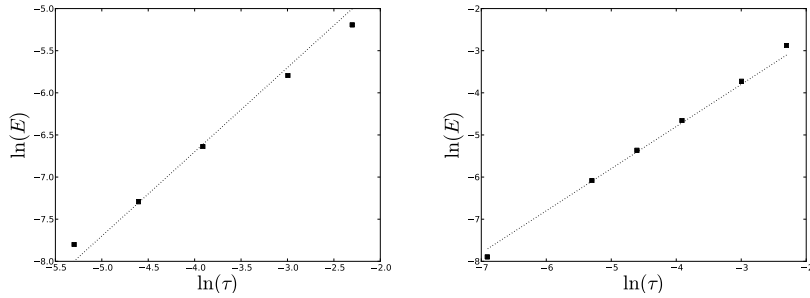
### Experiment 1

If we take the advection term $\boldsymbol{b}(u)$ to be zero and consider the nonlinearity $\theta(u) = u^{1/\gamma}$, where $\gamma > 1$, then the equation in the domain simplifies to the well-known porous medium equation. Providing appropriate initial and boundary conditions, a solution to this equation is the so-called Barenblatt solution (see for instance [45, Chapter 4.2.2])

$$u(\mathbf{x}, t) = \frac{1}{t^a}\left(c - \frac{\gamma - 1}{2\gamma} b \frac{|\mathbf{x}|^2}{t^{2b}}\right)_+^{\gamma/(\gamma-1)} \qquad (\mathbf{x} \in \mathbb{R}^d, t > 0),$$

where $a = \frac{d}{d(\gamma-1)+2}\,, b = a/d\,, c > 0$ and $u_+ = \max\{u, 0\}$. We take $d = 2$ and $\gamma = 2$ in particular. For this value of the exponent $\gamma$ the equation is also known as Boussinesq's equation . The function $\beta(u)$ will equal $u^{1/2}$. Assuming that this is the exact solution of our problem, we compute the function $g$ and modify the Dirichlet and the Neumann conditions in accordance with it. We solve the numerical problem on the time interval

$(0.1, 1.1)$ to avoid the singularity of the Barenblatt solution at $t = 0$. Figure 2.4(a) shows the dependence of the absolute error between the exact and numerical solution with respect to the time step $\tau$ in log-log scale. The dotted line has the slope 1.




(a) Barenblatt solution (b) Piecewise linear solution

Figure 2.4: Absolute errors with respect to the time step $\tau$

## Experiment 2

In this experiment we assume a linear advective vector field $\boldsymbol{b}(u) = (u, u, u)^T$. The nonlinear functions are $\theta(u) = u^{1/2}$ and $\beta(u) = u^2$. We test our scheme with the following exact solution.

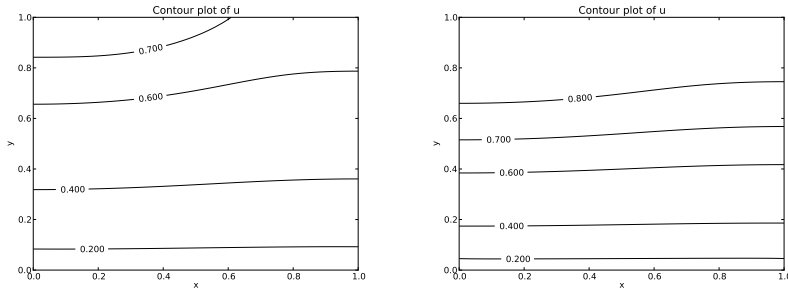
$$u(\mathbf{x}, t) = (y + 0.1x - 1 + t)_+^2$$

for which we compute $f$ and $g$. We solve the numerical problem on the time interval $(0.0, 1.0)$. Figure 2.4(b) shows the dependence of the absolute with respect to the time step $\tau$ in the log-log scale. We see that the behaviour of the error is similar as the one in the previous experiment.

## Experiment 3

We consider $\theta(u) = u^{1/2}$, $\beta(u) = u^{3/2}$, $\boldsymbol{b}(u) = \mathbf{0}$ and the zero source inside the domain ($f = 0$). The starting time is $T_0 = 0$ and the end time $T_{\text{end}} = 1.0$. We compare the numerical solution of (2.3) ("dynamical problem") to the one just with the Neumann boundary condition on $\Gamma_T$, i.e. the case when $\beta(u) \equiv 0$ ("Neumann problem"). Let us first consider time-constant boundary source $g(\mathbf{x}) = 0.5 + 0.2\sin(2\pi x)$. Figure 2.5 shows contour plots of the solution with dynamical and the Neumann boundary condition in two different times. The dynamical condition causes a certain delay of the

Figure 2.5: Contour plot of the numerical solutions for time-constant $g$ at $t = 1.0$

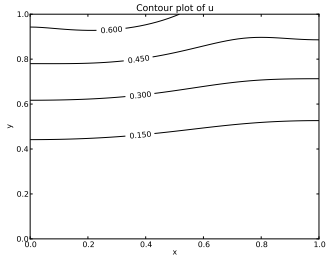



(a) Solution of "dynamical problem"

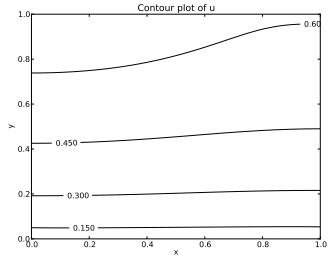
(b) Solution of "Neumann problem"

flow from the boundary. The time-dependent source case is more interesting. Figure 2.6 shows the results for $g(\mathbf{x}, t) = \Big(0.03 + 0.012\sin(2\pi x)\Big)/\Big(0.01 + (t - 0.5)^2\Big)$ in two different time. The function $g$ has stationary point in $t = 0.5$ as a function of time. We can here see that a saturation effect takes place and slows down the diffusion through the upper boundary. When $g$ starts decreasing in time, the delay is compensated and both solutions line up.

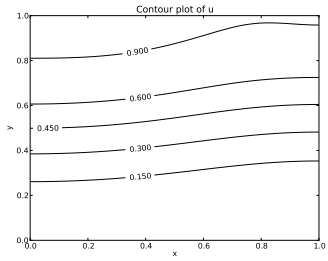
Figure 2.6: Contour plot of the numerical solutions for time-dependent $g$



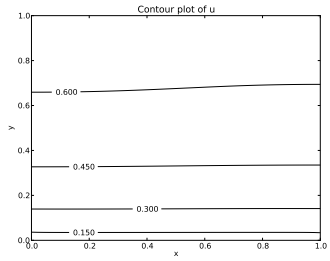

(a) Solution of "dynamical problem" at time $t = 0.5$




(b) Solution of "dynamical problem" at time $t = 1.0$




(c) Solution of "Neumann problem" at time $t = 0.5$




(d) Solution of "Neumann problem" at time $t = 1.0$

# Chapter 3

# On a nonlinear parabolic equation with a general boundary condition

This chapter treats a mathematical generalization of the boundary condition discussed in Chapter 2, which have a different physical background. We first put the problem under consideration into a wider context of standard and nonstandard boundary conditions for parabolic problems.

## 3.1 General boundary conditions

An unifying perspective on different boundary conditions can be found in the article [54]. It offers a derivation and physical interpretation of general boundary conditions on the example of the heat equation, which we will briefly resume.

The standard derivations of the heat equation are always based on the idea that "heat in equals heat out". Suppose we consider the heat flow in the region $\Omega \subseteq \mathbb{R}^d$ with smooth boundary $\partial\Omega$. Let $\boldsymbol{h}$ be the heat flow vector and $\boldsymbol{n}$ be the outward unit normal $\boldsymbol{n}$. Then the amount of heat flowing out of the region is

$$\int_{\partial\Omega} \boldsymbol{h} \cdot \boldsymbol{n} \, \mathrm{d}s.$$

Denote by $q$ the heat per unit volume. Conservation of heat, when phrased in integral form, says

$$\frac{d}{dt} \int_{\Omega} q \, \mathrm{d}x = - \int_{\partial\Omega} \boldsymbol{h} \cdot \boldsymbol{n} \, \mathrm{d}s + \int_{\Omega} s \, \mathrm{d}x. \tag{3.1}$$

The term on the left-hand side represents the change in heat content in $\Omega$ per unit time, which must be equal to the flux of heat through the boundary plus the contribution by the heat source $s$ in the region $\Omega$. The above equation becomes

$$\int_\Omega q_t \,\mathrm{d}x = -\int_\Omega \nabla \cdot \boldsymbol{h} \,\mathrm{d}x + \int_\Omega s \,\mathrm{d}x,$$

by interchanging the derivative and integral sign on the left-hand side, and by applying the divergence theorem for the boundary integral.

The heat content $q$ and the heat flow $\boldsymbol{h}$ depend on the temperature $u$ via the constitutive laws

$$q = \rho c u, \quad \text{and} \quad \boldsymbol{h} = -k\nabla u, \tag{3.2}$$

where $\rho$ is the density, $c$ is the heat capacity and $k$ is the thermal conductivity of the material. The latter formula is known as Fourier's law. The minus sign indicates that the heat flows from areas of higher temperature towards those with lower temperature. The constitutive laws, in general, reflect a relation between physical quantities that is specific to a material. Some of them are simply phenomenological and so they have to be often "verified" experimentally. From mathematical modelling point of view, constitutive laws can introduce nonlinearities into the equations.

One can combine (3.2) and the fact that (3.1) holds for an arbitrary subregion of $\Omega$ to rewrite the conservation of heat in its differential form

$$(\rho c u)_t = \nabla \cdot (k \nabla u) + s.$$

If the functions $\rho, c$ and $k$ are constants, the preceding takes a familiar form of the heat equation with the source $f$

$$\partial_t u - \alpha \Delta u = f, \tag{3.3}$$

where $\alpha = k/(\rho c)$ and $f = s/(\rho c)$. If any of the aforementioned functions is a nonlinear function of $u$, one speaks about the nonlinear heat diffusion.

In the traditional approach an equation like (3.3) is assumed to hold in the region $\Omega$ and the boundary conditions are appended later. There are three standard boundary conditions. The Dirichlet boundary condition specifies the temperature on the boundary

$$u(x,t) = u_{\partial\Omega}(x) \quad \text{for } x \in \partial\Omega \text{ and } t > 0.$$

The Neumann boundary condition specifies the heat flux on the boundary

$$\alpha \frac{\partial u}{\partial n}(x,t) = g(x) \quad \text{for } x \in \partial\Omega \text{ and } t > 0.$$

The third kind of the standard boundary conditions is the Robin boundary condition

$$\beta u(x,t) + \alpha \frac{\partial u}{\partial n}(x,t) = g(x) \quad \text{for } x \in \partial\Omega \text{ and } t > 0, \quad \beta > 0.$$

It can be used to incorporate Newton's law of cooling stating that the flux through the boundary between two regions is proportional to the temperature difference between them, that is

$$-\alpha \frac{\partial u}{\partial n}(x,t) = \beta(u - u_{\text{out}}).$$

It would be more natural to derive (general) boundary conditions in context of energy balance (3.1). Note that the function $g$ in both Neumann and Robin boundary condition can be seen as a heat source on the boundary. Let us state for this purpose the heat conservation law for the boundary

$$\frac{d}{dt}\int_{\partial\Omega} q \,\mathrm{d}s = -\int_{\partial\Omega} \boldsymbol{h}\cdot\boldsymbol{n}\,\mathrm{d}s + \int_{\partial\Omega}\phi\,\mathrm{d}s. \tag{3.4}$$

The term $\phi$ will represent a heat source on the boundary. It will also bear all the information about energy exchange and other interactions of the region $\Omega$ with the regions outside its boundary $\partial\Omega$. A sufficient condition for (3.4) to hold true is evidently

$$\partial_t q = -\boldsymbol{h}\cdot\boldsymbol{n} + \phi.$$

The constitutive equations (3.2) give in the simplest case

$$\partial_t u = -\alpha\nabla u\cdot\boldsymbol{n} + \widetilde{\phi}.$$

The left-hand side terms can be neglected for instance if $u = u(x)$ on $\partial\Omega$ or the contribution of $\partial_t q$ is negligible in comparision to the right-hand side. We recover in this way the Neumann boundary condition by setting $\widetilde{\phi} = g$ and the Robin boundary condition by $\widetilde{\phi} = g - \beta u$.

The term $\phi$ in (3.4) can be a function of the derivatives restricted to the boundary too. This leads to the so-called surface gradient and Laplace-Beltrami operator which acts in the directions tangential to $\partial\Omega$ and may allow for heat flow along the boundary. Their definitions are discussed in the next section. A very simple general boundary condition incorporating the Laplace-Beltrami operator reads as

$$\partial_t u = -\alpha\nabla u\cdot\boldsymbol{n} - \beta(u - u_{\text{out}}) + \Delta_\Gamma u \qquad \text{on } \partial\Omega \text{ and } t > 0.$$

It says that the time change of temperature $u$ on the boundary depends on the heat flux across the boundary, the temperature difference between the boundary and the outside region, and the heat diffusion flow along the boundary term (compare with [54]).

A different approach to derive general boundary conditions lies in applying an asymptotic analysis to a thin layer surrounding the domain (see[92] and [91]. The thin layer is consequently replaced by an approximating boundary condition. This approach also gives a valuable insight into the use of nonstandard boundary conditions in mathematical modelling.

A number of authors have focused in recent years on this kind of nonstandard boundary conditions (see [34, 116, 51]), but little attention has been paid to nonlinear diffusion. A semi-linear case is discussed in the paper [35]. The article [52] deals with a quasi-linear parabolic (possibly, degenerate) equations. The authors of [25] studied a system of parabolic equations with nonlinear coupling and dynamic boundary conditions. The reader is also referred to the papers [13] and [112] for interesting applications of the above-mentioned boundary conditions in liquid crystals and protocells respectively.

In this chapter we study a nonlinear degenerate parabolic equation accompanied by the nonlinear dynamical boundary condition of reactive-diffusive type. We namely consider the following initial-boundary value problem:

$$\begin{aligned} \partial_t \theta(u) - \Delta u &= f && \text{in } \Omega \times (0,T), \\ \partial_t \beta(u) - \Delta_\Gamma u + u + \nabla u \cdot \boldsymbol{n} &= 0 && \text{on } \Gamma \times (0,T), \\ u(x,0) &= u_0(x) && \text{in } \Omega \text{ and } \Gamma. \end{aligned} \tag{3.5}$$

The domain $\Omega$ is a sufficiently regular bounded domain in $\mathbb{R}^d, d \geq 1$ with the boundary $\partial\Omega = \Gamma$ and the outward normal vector $\boldsymbol{n}$. The data $f$ and $u_0$ are given. Both nonlinear functions $\theta, \beta : \mathbb{R} \to \mathbb{R}$ are continuous and almost everywhere differentiable.

We assume that the function $\theta$ is monotone and increasing ($0 \leq \theta'$), which can allow for the degeneracy of the problem (3.5). The function $\beta$ is strictly monotone ($0 < \lambda \leq \beta'$). They satisfy the linear growth condition

$$|\theta(s)| \leq C(1+|s|), \qquad |\beta(s)| \leq C(1+|s|),$$

and $\theta(0) = \beta(0) = 0$. Later, in the error estimates we will moreover assume the Lipschitz continuity of $\theta$ and $\beta$. There is already an extensive literature on degenerate parabolic equations and their numerical integration [39], [88]. We refer to [93] for error estimates of degenerate parabolic equations.

The aim of this chapter is to prove existence and uniqueness of a solution to (3.5), and demonstrate error estimates of time and space discretization.

## 3.2 Functional setting

The Laplace-Beltrami is defined as the divergence of the gradient of a function on a Riemannian manifold. We state a formal definition of the Laplace-Beltrami operator, but first we need to recall a few notions from differential geometry. For a comprehensive treatment of differential geometry, we refer the interested reader to the excellent and thorough book by Fecko [49].

The boundary $\Gamma$ can be seen as an oriented Riemannian manifold with the natural metric tensor $g$ inherited from $\mathbb{R}^d$. The metric tensor $g$ is given in local coordinates by the nonsingular symmetric coefficient matrix $(g_{ij})_{i,j=1,\dots,d-1}$. Let $|g| = \det(g_{ij})$ be the determinant of $g_{ij}$ and $g^{ij} = (g_{ij})^{-1}$ be the inverse matrix.

The gradient of a function $u$ on the manifold $\Gamma$ is the vector field $\mathrm{grad}_\Gamma u$ such that

$$\mathrm{grad}_\Gamma u \equiv \nabla_\Gamma u = \sum_{i,j=1}^{d-1} g^{ij} \frac{\partial u}{\partial x_j} \frac{\partial}{\partial x_i},$$

where the $\frac{\partial}{\partial x_i}$'s represent basis vectors pointing in the local coordinate directions. We write $H^1(\Gamma)$ for the Sobolev space of square-integrable functions with square integrable gradient on the manifold $\Gamma$. It is endowed with the norm

$$\|u\|_{H^1(\Gamma)} = \sqrt{\|u\|_\Gamma^2 + \|\nabla_\Gamma u\|_\Gamma^2}, \quad u \in H^1(\Gamma).$$

The reader is reminded that the subscript $\Gamma$ marks the integration over the boundary $\Gamma$

$$(u, v)_\Gamma = \int_\Gamma uv \,\mathrm{d}s, \quad \|u\|_\Gamma = \sqrt{(u, u)_\Gamma}.$$

The notation for domain integration is

$$(u, v) = \int_\Omega uv \,\mathrm{d}x, \quad \|u\| = \sqrt{(u, u)}.$$

The divergence of a vector field $F = F^i \frac{\partial}{\partial x_i}$ on the manifold $\Gamma$ is the function $\mathrm{div}_\Gamma F$ defined as the Lie derivative of the volume form on $\Gamma$ along the vector field $F$ (see [49, Section 8.2]). Its formula in local coordinates reads as

$$\mathrm{div}_\Gamma F = \sum_{i,j=1}^{d-1} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_i} \left( \sqrt{|g|} F^i \right).$$

The Laplace-Beltrami operator $\Delta_\Gamma$ applied to a function $u$ is in local form

$$\Delta_\Gamma u = \mathrm{div}_\Gamma \left(\mathrm{grad}_\Gamma u\right) = \frac{1}{\sqrt{|g|}} \sum_{i,j=1}^{d-1} \frac{\partial}{\partial x_i} \left( \sqrt{|g|} g^{ij} \frac{\partial u}{\partial x_j} \right).$$

We note that one can also define Laplace-Beltrami operator by differential forms.

As an basic example, let $\Omega = \{x \in \mathbb{R}^2 : |x| < R\}$ and $\Gamma = \partial\Omega$. The Laplace-Beltrami operator $\Delta_\Gamma$ on a function $u$ is in the polar coordinates $(r, \theta)$ given by

$$\Delta_\Gamma u = \frac{1}{R^2} \frac{\partial^2 u}{\partial \theta^2}.$$

We see that the operator $\Delta_\Gamma$ acts only in the directions tangential to $\Gamma$.

The Laplace-Beltrami operator can be defined for any $u$ from the Sobolev space $H^1(\Gamma)$. Since the manifold $\Gamma$ has no boundary ($\partial\Gamma = \emptyset$), the formula

$$-\int_\Gamma (\Delta_\Gamma u) v \,\mathrm{d}s = \int_\Gamma \nabla_\Gamma u \cdot \nabla_\Gamma v \,\mathrm{d}s \tag{3.6}$$

holds true for any smooth functions $u, v \in C^\infty(\Gamma)$. The density argument extends the above formula to any $u, v \in H^1(\Gamma)$. The integral on the left-hand side is then understood in the distributional sense as $\Delta_\Gamma u \in H^{-1}(\Gamma)$. More details on Laplace-Beltrami operator, Sobolev spaces on Riemannian manifolds and geometric analysis can be found in the book [68].

Taking the above considerations into account, we will work in the space

$$V = \{u \in H^1(\Omega) : u_{|\Gamma} \in H^1(\Gamma)\}$$

endowed with the graph norm

$$\|u\|_V = \sqrt{\|u\|^2_{H^1(\Omega)} + \|u\|^2_{H^1(\Gamma)}}.$$

The space $H^1(\Omega)$ is the standard Sobolev space of square-integrable functions on $\Omega$ with first weak derivatives. We use the Friedrichs inequality throughout the chapter in the version (see [94, Chapter 18] for instance)

$$\|\varphi\|^2 \le C\big(\,\|\nabla\varphi\|^2 + \|\varphi\|^2_\Gamma\,\big)$$

to make use of an equivalent norm on $H^1(\Omega)$.

A weak solution $u$ of the problem (3.5) accordingly satisfies

$$\begin{aligned}(\partial_t\theta(u),\varphi) + (\nabla u, \nabla\varphi) + (\partial_t\beta(u) + u, \varphi)_\Gamma + (\nabla_\Gamma u, \nabla_\Gamma\varphi)_\Gamma = (f,\varphi)\,,\\ u(0) = u_0\end{aligned} \tag{3.7}$$

for every $\varphi \in V$ and a.e. $t \in (0,T)$.

**Remark 3.1.** *If the general boundary condition in (3.5) is imposed just on a part of the boundary (i. e. $\Gamma \subsetneq \partial\Omega$), the situation becomes interesting. A boundary term appears in (3.6) unlike in our case, because $\partial\Gamma \neq \emptyset$. This term has to be then handled by prescribing an appropriate boundary condition for it.*

## 3.3 Time discretization

In this section we prove the unique solvability of the problem (3.7) and establish error estimates for the time discretization scheme. We follow the already introduced notation for a time-discretized function and its forward difference

$$w(t_i) = w_i, \qquad \delta w_i = \frac{w_i - w_{i-1}}{\tau},$$

where the time $t_i = i\tau$, the uniform time step $\tau = T/n$ and $n \in \mathbb{N}$.

A time-discretized approximation of the problem (3.7) reads as

$$\begin{aligned}(\delta\theta(u_i), \varphi) + (\nabla u_i, \nabla\varphi) + (\delta\beta(u_i) + u_i, \varphi)_\Gamma \\ + (\nabla_\Gamma u_i, \nabla_\Gamma \varphi)_\Gamma = (f_i, \varphi)\,, \qquad \forall \varphi \in V \text{ and } i = 1, \dots, n.\end{aligned} \tag{3.8}$$

The theory of monotone operators guarantees that for the every index $i$ there exists a unique solution.

**Lemma 3.1.** *Assume $u_0 \in V$, then there exists a uniquely determined $u_i \in V$ solving (3.8) for any index $i = 1, \dots, n$.*

*Proof.* Define first the operator $A$

$$A : V \to V^*, \qquad u \mapsto A(u)$$

which is given by the formula

$$\begin{aligned}\langle A(u), \varphi\rangle = &(\theta(u), \varphi) + \tau\,(\nabla u, \nabla\varphi) \\ &+ (\beta(u) + \tau u, \varphi)_\Gamma + \tau\,(\nabla_\Gamma u, \nabla_\Gamma\varphi)_\Gamma\,.\end{aligned}$$

We can easily deduce that the operator $A$ is strictly monotone

$$\begin{aligned}\langle A(u) - A(v), u - v\rangle = &(\theta(u) - \theta(v), u - v) + \tau \left\|\nabla(u - v)\right\|^2 \\ &+ (\beta(u) - \beta(v), u - v)_\Gamma + \tau \left\|u - v\right\|_\Gamma^2 + \tau \left\|\nabla_\Gamma(u - v)\right\|_\Gamma^2 \\ \geq &\tau \left\|\nabla(u - v)\right\|^2 + (\lambda + \tau) \left\|u - v\right\|_\Gamma^2 + \tau \left\|\nabla(u - v)\right\|_\Gamma^2 \\ \geq &C \left\|u - v\right\|_V^2\,.\end{aligned}$$

It is also hemicontinuous and coercive

$$\begin{aligned}\frac{\langle A(u), u\rangle}{\left\|u\right\|_V} &\geq \frac{\tau \left\|\nabla u\right\|^2 + (\lambda + \tau) \left\|u\right\|_\Gamma^2 + \tau \left\|\nabla u\right\|_\Gamma^2}{\left\|u\right\|_V} \\ &\geq C \left\|u\right\|_V \to \infty \quad \text{ for } \left\|u\right\|_V \to \infty.\end{aligned}$$

Theorem A.10 can be consequently invoked. □

We use Rothe's method, which provides a suitable functional framework to prove the unique solvability of the original problem (3.7). The notation for the Rothe functions remains the same as in the previous chapter. We set up the piecewise-constant-in-time function $\overline{u}_n$

$$\overline{u}_n(0) = u(0), \qquad \overline{u}_n(t) = u_i \qquad \text{for } t \in (t_{i-1}, t_i],$$

and the piecewise-linear-in-time function $u_n$

$$u_n(0) = u(0), \qquad u_n(t) = u_{i-1} + \delta u_i (t - t_{i-1}) \qquad \text{for } t \in (t_{i-1}, t_i].$$

The piecewise-linear-in-time functions $\theta_n$ and $\beta_n$ are defined likewise

$$\begin{aligned} \theta_n(0) &= \theta(u(0)), \qquad \theta_n(t) = \theta(u_{i-1}) + \delta\theta(u_i)(t - t_{i-1}) \qquad \text{for } t \in (t_{i-1}, t_i], \\ \beta_n(0) &= \beta(u(0)), \qquad \beta_n(t) = \beta(u_{i-1}) + \delta\beta(u_i)(t - t_{i-1}) \qquad \text{for } t \in (t_{i-1}, t_i]. \end{aligned}$$

The sequence of BVPs (3.8) is consequently replaced by one problem which reads as

$$\begin{aligned} &(\partial_t \theta_n, \varphi) + (\nabla \overline{u}_n, \nabla \varphi) + (\partial_t \beta_n + \overline{u}_n, \varphi)_\Gamma \\ &+ (\nabla_\Gamma \overline{u}_n, \nabla_\Gamma \varphi)_\Gamma = \left(\overline{f}_n, \varphi\right) \qquad \forall \varphi \in V. \end{aligned} \tag{3.9}$$

The next lemma concerns a priori estimates, which will provide the essential information about the solution.

**Lemma 3.2.** *Assume $f' \in L^2((0,T), L^2(\Omega))$ and $u_0 \in L^2(\Omega)$. If $u_i$ is the solution of (3.8), then there exists a positive constant $C$ such that*

*(i)*

$$\sum_{i=1}^{n} \left( \|\nabla u_i\|^2 + \|u_i\|_\Gamma^2 + \|\nabla_\Gamma u_i\|_\Gamma^2 \right) \tau \le C,$$

*(ii)*

$$\begin{aligned} &\max_{1 \le j \le n} \left( \|\nabla u_j\|^2 + \|u_j\|_\Gamma^2 + \|\nabla_\Gamma u_j\|_\Gamma^2 \right) + \sum_{i=1}^{n} \tau \|\delta u_i\|_\Gamma^2 \\ &+ \sum_{i=1}^{n} \left( \|\nabla (u_i - u_{i-1})\|^2 + \|u_i - u_{i-1}\|_\Gamma^2 + \|\nabla_\Gamma (u_i - u_{i-1})\|_\Gamma^2 \right) \le C. \end{aligned}$$

*Proof.* The proof will go along the lines of the proofs of Lemma (2.2) and (2.3) respectively.

$(i)$ Putting $\varphi = \tau u_i$ in (3.8) and adding it up for $i = 1, \dots, n$ yields

$$\begin{aligned} &\sum_{i=1}^{n} (\delta\theta(u_i), u_i)\, \tau + \sum_{i=1}^{n} \|\nabla u_i\|^2 \tau + \sum_{i=1}^{n} (\delta\beta(u_i), u_i)_\Gamma \tau \\ &+ \sum_{i=1}^{n} \|u_i\|_\Gamma \tau + \sum_{i=1}^{n} \|\nabla_\Gamma u_i\|_\Gamma^2 \tau = \sum_{i=1}^{n} (f_i, u_i)\, \tau. \end{aligned} \tag{3.10}$$

We recall the inequality (2.12) from the second chapter to bound the first sum

$$\begin{aligned}
&\sum_{i=1}^{n} (\theta(u_i) - \theta(u_{i-1}), u_i) \\
&= (\theta(u_n), u_n) - (\theta(u_0), u_0) - \sum_{i=1}^{n} (u_i - u_{i-1}, \theta(u_{i-1})) \\
&\geq (\theta(u_n), u_n) - (\theta(u_0), u_0) - \sum_{i=1}^{n} \int_\Omega \big(\Phi_\theta(u_i) - \Phi_\theta(u_{i-1})\big) \\
&= \left[(\theta(u_n), u_n) - \int_\Omega \Phi_\theta(u_n)\right] - \left[(\theta(u_0), u_0) - \int_\Omega \Phi_\theta(u_0)\right] \\
&= \int_\Omega \left[\tilde{\Phi}_\theta(u_n) - \tilde{\Phi}_\theta(u_0)\right] \\
&\geq -(\theta(u_0), u_0) \\
&\geq -C.
\end{aligned}$$

The same reasoning applies also to the term containing $\beta$ function

$$\sum_{i=1}^{n} (\delta\beta(u_i), u_i)_{\Gamma_T} \tau \geq -C.$$

The $\varepsilon$-Young and then the Friedrich inequality imply for the right-hand side of (3.10) that

$$\sum_{i=1}^{n} (f_i, u_i)\, \tau \leq \sum_{i=1}^{n} \big(C_\varepsilon \|f_i\|^2 + \varepsilon \|u_i\|^2\big)\tau \leq C_\varepsilon + \varepsilon \sum_{i=1}^{n} \big( \|\nabla u_i\|^2 + \|u_i\|_\Gamma^2 \big)\tau.$$

$(ii)$ Putting $\varphi = \tau\delta u_i$ this time in (3.8) and adding it up for $i = 1, \ldots, j$ gives us

$$\begin{aligned}
&\sum_{i=1}^{j} (\delta\theta(u_i), \delta u_i)\, \tau + \sum_{i=1}^{j} (\nabla u_i, \nabla\delta u_i)\, \tau + \sum_{i=1}^{j} (\delta\beta(u_i), \delta u_i)_\Gamma\, \tau \\
&+ \sum_{i=1}^{j} (u_i, \delta u_i)_\Gamma\, \tau + \sum_{i=1}^{j} (\nabla_\Gamma u_i, \nabla_\Gamma \delta u_i)_\Gamma\, \tau = \sum_{i=1}^{j} (f_i, \delta u_i)\, \tau.
\end{aligned} \tag{3.11}$$

Since the functions $\theta$ and $\beta$ are monotone, it follows that

$$\sum_{i=1}^{j} (\delta\theta(u_i), \delta u_i)\, \tau + \sum_{i=1}^{j} (\delta\beta(u_i), \delta u_i)_\Gamma\, \tau \geq \sum_{i=1}^{j} \lambda \|\delta u_i\|_\Gamma^2\, \tau.$$

The remaining terms on the LHS of (3.11) can be rewritten by the Abel summation, e.g.

$$\sum_{i=1}^{j} (\nabla_\Gamma u_i, \nabla_\Gamma \delta u_i)\,\tau = \frac{1}{2}\Big( \|\nabla_\Gamma u_j\|^2 - \|\nabla_\Gamma u_0\|^2 + \sum_{i=1}^{j} \|\nabla_\Gamma (u_i - u_{i-1})\|^2 \Big).$$

We observe for the RHS of (3.11), as in the proof of Lemma (2.3), that

$$\begin{aligned}
\sum_{i=1}^{j} (f_i, \delta u_i)\,\tau &= (f_j, u_j) - (f_0, u_0) - \sum_{i=1}^{j} (\delta f_i, u_{i-1})\,\tau \\
&\le C_\varepsilon \|f_j\|^2 + \varepsilon \|u_j\|^2 + C + C \sum_{i=1}^{j} \Big( \|\delta f_i\|^2 + \|u_{i-1}\|^2 \Big)\tau \\
&\le \varepsilon C\big( \|u_j\|_\Gamma^2 + \|\nabla u_j\|^2 \big) + C_\varepsilon .
\end{aligned}$$

After choosing a sufficiently small $\varepsilon > 0$, we conclude the proof by taking maximum over $j = 1, \dots, n$. □

We are now in the position to prove a basic result of this chapter.

**Theorem 3.1.** *Suppose that $u_0 \in V$ and $f' \in L^2((0,T), L^2(\Omega))$. Then there exists the unique solution of (3.7).*

*Proof.* We demonstrate that the Rothe sequences of approximate solutions and $u_n$ converge to a limit $u$, which is a solution of (3.7). Lemma 3.2 $(i)$ guarantees that the sequence $\overline{u}_n$ is bounded in the reflexive space $L^2((0,T), V)$, more exactly

$$\begin{aligned}
\|\overline{u}_n\|^2_{L^2((0,T),V)} &= \int_0^T \|\overline{u}_n\|_V^2 \,\mathrm{d}t \\
&\le \int_0^T C\big( \|\nabla \overline{u}_n\|^2 + \|\overline{u}_n\|_\Gamma^2 + \|\nabla_\Gamma \overline{u}_n\|_\Gamma^2 \big)\,\mathrm{d}t \\
&= \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} C\big( \|u_i\|^2 + \|u_i\|_\Gamma^2 + \|\nabla_\Gamma u_i\|_\Gamma^2 \big)\,\mathrm{d}t \\
&= C \sum_{i=1}^{n} \big( \|u_i\|^2 + \|u_i\|_\Gamma^2 + \|\nabla_\Gamma u_i\|_\Gamma^2 \big)\tau \\
&\le C.
\end{aligned} \tag{3.12}$$

We can consequently select a subsequence $\mathbb{N}' \subset \mathbb{N}$ such that

$$\overline{u}_n \rightharpoonup u \qquad \text{in } L^2((0,T), V) \tag{3.13}$$

for $n \in \mathbb{N}'$. The sequence $u_n$ converges to the same $u$ according to Lemma 3.2 $(ii)$

$$\begin{aligned}\int_0^T \|\overline{u}_n - u_n\|_V^2 \,\mathrm{d}t &= \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|(\tau - t + t_{i-1})\delta u_i\|_V^2 \,\mathrm{d}t \\ &\le \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|2(u_i - u_{i-1})\|_V^2 \,\mathrm{d}t \\ &\le C \sum_{i=1}^n \Big( \|u_i - u_{i-1}\|^2 + \|u_i - u_{i-1}\|_\Gamma^2 \\ &\qquad + \|\nabla_\Gamma (u_i - u_{i-1})\|_\Gamma^2 \Big) \tau \\ &\le C\tau.\end{aligned}$$

We further claim for every $n \in \mathbb{N}'$ that

$$\int_0^T \|\overline{u}_n(t+s) - \overline{u}_n(t)\|^2 \,\mathrm{d}t \le Cs \qquad \text{for any small } s \in \mathbb{R}. \tag{3.14}$$

If the point $t + s$ lies outside the interval $(0, T)$, the time function $\overline{u}_n$ is additionally extended as a constant. To confirm (3.14), we start by the following observation for the distributional derivative $\partial_t \overline{u}_n$

$$\begin{aligned}\int_0^T \|\partial_t \overline{u}_n\|_V^2 \,\mathrm{d}t &= \int_0^T \sum_{i=1}^n \|u_i - u_{i-1}\|_V^2 \,\delta_{t_i}(t) \,\mathrm{d}t \\ &= \sum_{i=1}^n \|u_i - u_{i-1}\|_V^2 \\ &\le \sum_{i=1}^n C\big( \|\nabla(u_i - u_{i-1})\|^2 + \|u_i - u_{i-1}\|_\Gamma^2 + \|\nabla_\Gamma (u_i - u_{i-1})\|_\Gamma^2 \big) \\ &\le C.\end{aligned}$$

Here, the symbol $\delta_{t_i}$ denotes the Dirac delta function at the point $t_i$ which can be regarded as a discrete measure, i.e.

$$\int_0^T \phi(t)\delta_{t_i}(t) \,\mathrm{d}t = \int_0^T \phi(t) \,\mathrm{d}\delta_{t_i} = \phi(t_i).$$

The application of Cauchy's inequality and the above estimate establishes the claim

$$\begin{aligned}\int_0^T \|\overline{u}_n(t+s)-\overline{u}_n(t)\|_V^2 \,\mathrm{d}t &= \int_0^T \left\| \int_0^1 \partial_t \overline{u}_n(t+\vartheta s) s \,\mathrm{d}\vartheta \right\|_V^2 \mathrm{d}t \\ &\le \int_0^T \int_0^1 \|\partial_t \overline{u}_n(t+\vartheta s) s\|_V^2 \,\mathrm{d}\vartheta \,\mathrm{d}t \\ &\le \int_0^T \|\partial_t \overline{u}_n(t+\vartheta s)\|_V^2 \, s^2 \,\mathrm{d}t \\ &\le C s^2.\end{aligned}$$

The similar result holds true for the space variable too. If the norm of the difference $h \in \mathbb{R}^d$ is sufficiently small, we have the bound for the domain integral

$$\begin{aligned}&\int_0^T \int_\Omega |\overline{u}_n(x+h,t)-\overline{u}_n(x,t)|^2 \,\mathrm{d}x \,\mathrm{d}t \\ \le &\int_0^T \int_\Omega |\nabla \overline{u}_n(x,t)|^2 h^2 \,\mathrm{d}x \,\mathrm{d}t \le C|h|^2.\end{aligned}$$

Supposing that $x+h \in \Gamma$, we can also derive that

$$\int_0^T \int_\Gamma |\overline{u}_n(x+h,t)-\overline{u}_n(x,t)|^2 \,\mathrm{d}s \,\mathrm{d}t \le C|h|^2.$$

These bounds, together with (3.12), (3.14), and Riesz-Kolmogorov theorem (see Theorem A.19) imply that the sequence $\overline{u}_n$ is relatively compact in the spaces $L^2(\Omega\times(0,T))$ and $L^2(\Gamma\times(0,T))$. We can a fortiori choose a subsequence converging almost everywhere

$$\overline{u}_n \to u \quad \text{a.e. in } \Omega\times(0,T) \text{ and } \Gamma\times(0,T). \tag{3.15}$$

Lemma 3.2 $(i)$ furthermore ensures that the time derivative $\partial_t u \in L^2((0,T),L^2(\Gamma))$

$$\int_0^T \|\partial_t u_n\|_\Gamma^2 \,\mathrm{d}t = \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|\partial_t u_n\|_\Gamma^2 \,\mathrm{d}t = \sum_{i=1}^n \|\delta u_i\|_\Gamma^2 \,\tau \le C.$$

The function $u$ thus belongs to the space of continuous functions $C((0,T),L^2(\Gamma))$.

The integration of (3.9) in time produces

$$\begin{aligned}&(\theta(\overline{u}_n(t)),\varphi) + (\beta(\overline{u}_n(t)),\varphi)_\Gamma + \langle D_n(t),\varphi\rangle \\ &-\big[\,(\theta(u_0),\varphi)+(\beta(u_0),\varphi)_\Gamma\,\big] + \int_0^t (\nabla\overline{u}_n,\nabla\varphi) \\ &+\int_0^t \Big[\,(\overline{u}_n,\varphi)_\Gamma + (\nabla_\Gamma \overline{u}_n,\nabla_\Gamma\varphi)_\Gamma\,\Big] = \int_0^t \big(\overline{f}_n,\varphi\big)\,,\end{aligned} \tag{3.16}$$

where we have introduced the functional $D_n(t)$ as follows

$$\langle D_n(t), \varphi \rangle = (\theta_n(t) - \theta(\overline{u}_n(t)), \varphi) + (\beta_n(t) - \beta(\overline{u}_n(t)), \varphi)_\Gamma .$$

The functional $D_n(t)$ vanishes as $n \to \infty$. Indeed, we deduce from Lemma 3.2 $(ii)$ for an arbitrary $t \in (t_{i-1}, t_i]$ that

$$\begin{aligned} \lim_{n\to\infty} |\langle D_n(t), \varphi \rangle| &= \lim_{n\to\infty} |t - t_i| \Big| (\partial_t \theta_n(t), \varphi) + (\partial_t \beta_n(t), \varphi)_\Gamma \Big| \\ &\le \lim_{n\to\infty} \tau \Big| \big(\overline{f}_n, \varphi\big) - (\nabla \overline{u}_n, \nabla \varphi) - (\overline{u}_n, \varphi)_\Gamma - (\nabla_\Gamma \overline{u}_n, \nabla_\Gamma \varphi)_\Gamma \Big| \\ &\le \lim_{n\to\infty} \frac{C \, \|\varphi\|_V}{n} \\ &= 0. \end{aligned} \tag{3.17}$$

We finally send $n \to \infty$ in (3.16). Combining the assertions (3.13), (3.15), (3.17) and the Lebesgue dominated theorem we discover that

$$\begin{aligned} &(\theta(u(t)), \varphi) + (\beta(u(t)), \varphi)_\Gamma - (\theta(u_0), \varphi) + (\beta(u_0), \varphi)_\Gamma \\ &+ \int_0^t \Big[ (\nabla u, \nabla \varphi) + (u, \varphi)_\Gamma + (\nabla_\Gamma u, \nabla_\Gamma \varphi)_\Gamma \Big] = \int_0^t (f, \varphi) . \end{aligned} \tag{3.18}$$

The first term in the LHS of (3.18) can be extended continuously for every $t \in [0, T]$, because all the other terms are well defined on the whole interval. The same applies for its differentiability with respect to time and so differentiation shows that $u$ is indeed a solution of (3.7).

What remains is to demonstrate the uniqueness of the solution. To do so, we assume for the sake of contradiction that both $u$ and $v$ satisfy (3.7). We subtract the corresponding identities from each other, integrate in time, set the test function $\varphi = u - v$ and integrate in time again to get

$$\begin{aligned} &\int_0^t (\theta(u) - \theta(v), u - v) + \int_0^t (\beta(u) - \beta(v), u - v)_\Gamma \\ &+ \int_0^t \left( \int_0^s \nabla(u - v), \nabla(u - v) \right) + \int_0^t \left( \int_0^s \nabla_\Gamma(u - v), \nabla_\Gamma(u - v) \right)_\Gamma \\ &+ \int_0^t \left( \int_0^s (u - v), u - v \right)_\Gamma = 0. \end{aligned} \tag{3.19}$$

The first two terms on the LHS of (3.19) are non-negative by the monotonicity argument. For the next term in (3.19) it is easy to verify by integration by parts that

$$\int_0^t \left( \int_0^s \nabla(u - v), \nabla(u - v) \right) = \frac{1}{2} \left\| \int_0^t \nabla(u - v) \right\|^2 \ge 0. \tag{3.20}$$

The same conclusion can be drawn for the last two integrals, which is a contradiction with zero right hand side of (3.19) unless $u = v$. □

The next theorem provides error estimates for the time discretization. To obtain better result for the approximate solution in the domain, we will assume Lipschitz continuity of the function $\theta$.

**Theorem 3.2.** *Suppose that $u_0 \in V$ and $f' \in L^2((0,T), L^2(\Omega))$.*

*(i) Then there exists a positive constant $C$ such that*

$$\int_0^T \|\overline{u}_n - u\|_\Gamma^2 \leq C\tau.$$

*(ii) If moreover $\theta$ is Lipschitz continuous with the Lipschitz constant $L$ (i.e $\theta' < L$), then there exists a positive constant $C$ such that*

$$\int_0^T \|\theta(\overline{u}_n) - \theta(u)\|^2 + \int_0^T \|\overline{u}_n - u\|_\Gamma^2 \leq C\tau.$$

*Proof.* $(i)$ We subtract (3.7) from (3.9) and integrate in time. We set $\varphi = \overline{u}_n - u$, integrate in time once more to see that

$$\begin{aligned}
&\int_0^t (\theta(\overline{u}_n) - \theta(u), \overline{u}_n - u) \\
&+ \int_0^t \left( \int_0^s \nabla(\overline{u}_n - u), \nabla(\overline{u}_n - u) \right) + \int_0^t (\beta(\overline{u}_n) - \beta(u), \overline{u}_n - u)_\Gamma \\
&+ \int_0^t \left( \int_0^s (\overline{u}_n - u), \overline{u}_n - u \right)_\Gamma + \int_0^t \left( \int_0^s \nabla(\overline{u}_n - u), \nabla(\overline{u}_n - u) \right)_\Gamma \\
&= \int_0^t \Big[ (\theta(\overline{u}_n) - \theta_n, \overline{u}_n - u) + (\beta(\overline{u}_n) - \beta_n, \overline{u}_n - u)_\Gamma \Big].
\end{aligned} \tag{3.21}$$

Since the function $\theta$ is monotone and $\beta$ strongly monotone, it follows that

$$\int_0^t (\theta(\overline{u}_n) - \theta(u), \overline{u}_n - u) + \int_0^t (\beta(\overline{u}_n) - \beta(u), \overline{u}_n - u)_\Gamma \geq \int_0^t \lambda \|\overline{u}_n - u\|_\Gamma^2 .$$

An observation equivalent to (3.20) holds for the next three terms. Utilizing (3.17) and (3.13) gives this upper bound for the RHS of (3.21)

$$\begin{aligned}
&\int_0^t \Big[ (\theta(\overline{u}_n) - \theta_n, \overline{u}_n - u) + (\beta(\overline{u}_n) - \beta_n, \overline{u}_n - u)_\Gamma \Big] \\
&= - \int_0^t \langle D_n(s), \overline{u}_n - u \rangle \leq C\tau \int_0^t \|\overline{u}_n - u\|_V \leq C\tau.
\end{aligned}$$

The proof of $(i)$ finishes by taking $t = T$.

$(ii)$ The proof runs as before except the $\theta$-term on the RHS of (3.21), where we write

$$\int_0^t (\theta(\overline{u}_n) - \theta(u), \overline{u}_n - u) \geq \int_0^t \frac{1}{L} \left\|\theta(\overline{u}_n) - \theta(u)\right\|^2 .$$

□

## 3.4 Full discretization

This section deals with the space discretization of (3.8), which completes the full discretization of the original problem (3.7). We also state the result concerning first-order Lagrange finite element space discretization. To begin with, let $V_h$ be a finite-dimensional subspace of $V$ and $\pi_h : V \to V_h$ a linear bounded projection onto it.

The fulldiscrete problem consists in finding the functions $u_i^h \in V_h$ such that

$$\begin{aligned} \left(\delta\theta_i^h, \varphi\right) + \left(\nabla u_i^h, \nabla\varphi\right) + \left(\delta\beta_i^h + u_i^h, \varphi\right)_\Gamma + \left(\nabla_\Gamma u_i^h, \nabla_\Gamma\varphi\right)_\Gamma \\ = \left(f_i^h, \varphi\right) \end{aligned} \tag{3.22}$$

is satisfied for any $\varphi \in V_h$ and $i = 1, \ldots, n$. In terms of the Rothe functions, it reads as

$$\begin{aligned} \left(\partial_t\theta_n^h, \varphi\right) + \left(\nabla\overline{u}_n^h, \nabla\varphi\right) + (\partial_t\beta_n^h + \overline{u}_n^h, \varphi)_\Gamma + \left(\nabla_\Gamma\overline{u}_n^h, \nabla_\Gamma\varphi\right)_\Gamma \\ = \left(\overline{f}_n^h, \varphi\right), \quad \forall\varphi \in V_h \\ u_n^h(0) = \pi_h u_0, \end{aligned} \tag{3.23}$$

which is just a space-discretized version of (3.9). The functions $\overline{u}_n^h, \theta_n^h, \beta_n^h$ and $\overline{f}_n^h$ represent the space-discretized counterparts of $\overline{u}_n^h, \theta_n, \beta_n$ and $\overline{f}_n$ respectively. The existence of the unique solution $u_n^h$ can be demonstrated by the same method as in the previous section. For this reason we omit the proof of the assertion below.

**Lemma 3.3.** *Suppose that $u_0 \in V$ and $f' \in L^2((0,T), L^2(\Omega))$. Then there exists a unique solution of (3.22).*

The next theorem provides a space discretization error estimate between the solution $u$ of (3.7) and the solution of $\overline{u}_n^h$ (3.22).

**Theorem 3.3.** *Suppose that $u_0 \in V$, $f' \in L^2((0,T), L^2(\Omega))$ and the functions $\theta$ and $\beta$*

*are Lipschitz continuous with the Lipschitz constant L. Then the error estimate*

$$
\begin{aligned}
&\int_0^t \left\|\theta(\overline{u}_n^h) - \theta(u)\right\|^2 + \int_0^t \left\|\overline{u}_n^h - u\right\|_\Gamma^2 \\
&+ \left\|\int_0^t \nabla(\overline{u}_n^h - u)\right\|^2 + \left\|\int_0^t (\overline{u}_n^h - u)\right\|_\Gamma^2 + \left\|\int_0^t \nabla(\overline{u}_n^h - u)\right\|_\Gamma^2 \\
&\leq C\Bigg(\tau + \|\theta(u_0) - \theta(\pi_h u_0)\| + \|\beta(u_0) - \beta(\pi_h u_0)\|_\Gamma \\
&+ \sqrt{\int_0^t \|u - \pi_h u\|^2} + \sqrt{\int_0^t \|u - \pi_h u\|_\Gamma^2} \\
&+ \int_0^t \left[\, \|\nabla(u - \pi_h u)\|^2 + \|u - \pi_h u\|_\Gamma^2 + \|\nabla_\Gamma(u - \pi_h u)\|_\Gamma^2 \,\right]\Bigg)
\end{aligned}
$$

*holds for any* $t \in (0, T)$.

*Proof.* The proof proceeds as in Theorem 3.2. In the similar manner, the subtraction of (3.7) from (3.23), subsequent integration in time, substitution $\varphi = \overline{u}_n^h - \pi_h u$ and repeated integration in time produce

$$
\begin{aligned}
&\int_0^t \left(\theta_n^h - \theta(u), \overline{u}_n^h - \pi_h u\right) \\
&+ \int_0^t \left(\int_0^s \nabla(\overline{u}_n^h - u), \nabla(\overline{u}_n^h - \pi_h u)\right) + \int_0^t \left(\beta_n^h - \beta(u), \overline{u}_n^h - \pi_h u\right)_\Gamma \\
&+ \int_0^t \left(\int_0^s \overline{u}_n^h - u, \overline{u}_n^h - \pi_h u\right)_\Gamma + \int_0^t \left(\int_0^s \nabla_\Gamma(\overline{u}_n^h - u), \nabla_\Gamma(\overline{u}_n^h - \pi_h u)\right)_\Gamma \\
&= \int_0^t \left(\theta(u_0) - \theta(\pi_h u_0), \overline{u}_n^h - \pi_h u\right) + \int_0^t \left(\beta(u_0) - \beta(\pi_h u_0), \overline{u}_n^h - \pi_h u\right)_\Gamma .
\end{aligned}
$$

We use the common trick of adding $\pm\theta(\overline{u}_n^h)$, $\pm\beta(\overline{u}_n^h)$ and $\pm u$ respectively to appropriate

terms to get the difference $u_n^h - u$ in the right hand:

$$
\begin{aligned}
&\int_0^t \big(\theta(u_n^h) - \theta(u), \overline{u}_n^h - u\big) + \frac{1}{2}\left\|\int_0^t \nabla(u_n^h - u)\right\|^2 \\
&+ \int_0^t \big(\beta(u_n^h) - \beta(u), \overline{u}_n^h - u\big)_\Gamma + \frac{1}{2}\left\|\int_0^t (u_n^h - u)\right\|_\Gamma^2 + \frac{1}{2}\left\|\int_0^t \nabla_\Gamma(u_n^h - u)\right\|_\Gamma^2 \\
&= \int_0^t \big(\theta(u_0) - \theta(\pi_h u_0), \overline{u}_n^h - \pi_h u\big) + \int_0^t \big(\beta(u_0) - \beta(\pi_h u_0), \overline{u}_n^h - \pi_h u\big)_\Gamma \qquad (3.24)\\
&- \int_0^t \Big[\big(\theta(\overline{u}_n^h) - \theta(u), u - \pi_h u\big) + \big(\beta(\overline{u}_n^h) - \beta(u), u - \pi_h u\big)_\Gamma\Big] \\
&- \int_0^t \Big[\big(\theta_n^h - \theta(\overline{u}_n^h), \overline{u}_n^h - \pi_h u\big) + \big(\beta_n^h - \beta(\overline{u}_n^h), \overline{u}_n^h - \pi_h u\big)_\Gamma\Big] \\
&- \int_0^t \left(\int_0^s \nabla(\overline{u}_n^h - u), \nabla(u - \pi_h u)\right) \\
&- \int_0^t \left(\int_0^s (\overline{u}_n^h - u), (u - \pi_h u)\right)_\Gamma - \int_0^t \left(\int_0^s \nabla_\Gamma(\overline{u}_n^h - u), \nabla_\Gamma(u - \pi_h u)\right)_\Gamma.
\end{aligned}
$$

The LHS of (3.24) is handled in much the same way as in the proof of Theorem 3.2. Let us now estimate the RHS of (3.24) step by step. Note that we have already the estimates for the solutions $\overline{u}_n^h$ and $u$ from the previous results. Hölder inequality for the first two terms coming from the initial condition yields

$$
\begin{aligned}
&\int_0^t \big(\theta(u_0) - \theta(\pi_h u_0), \overline{u}_n^h - \pi_h u\big) + \int_0^t \big(\beta(u_0) - \beta(\pi_h u_0), \overline{u}_n^h - \pi_h u\big)_\Gamma \\
\leq& C\big(\,\|\theta(u_0) - \theta(\pi_h u_0)\| + \|\beta(u_0) - \beta(\pi_h u_0)\|_\Gamma\,\big),
\end{aligned}
$$

where we have used the fact that

$$
\big\|\overline{u}_n^h - \pi_h u\big\|_\Gamma \leq \big\|\overline{u}_n^h - \pi_h u\big\| \leq \big\|\overline{u}_n^h\big\| + \|\pi_h u\| \leq C.
$$

Further, it follows from the growth conditions on $\theta$ and $\beta$ that

$$
\begin{aligned}
&- \int_0^t \Big[\big(\theta(\overline{u}_n^h) - \theta(u), u - \pi_h u\big) + \big(\beta(\overline{u}_n^h) - \beta(u), u - \pi_h u\big)_\Gamma\Big] \\
&\qquad\qquad \leq C\left(\sqrt{\int_0^t \|u - \pi_h u\|^2} + \sqrt{\int_0^t \|u - \pi_h u\|_\Gamma^2}\right).
\end{aligned}
$$

Reusing the argument (3.22) for the third line of the RHS in (3.24) implies

$$
- \int_0^t \Big[\big(\theta_n^h - \theta(\overline{u}_n^h), \overline{u}_n^h - \pi_h u\big) + \big(\beta_n^h - \beta(\overline{u}_n^h), \overline{u}_n^h - \pi_h u\big)_\Gamma\Big] \leq C\tau.
$$

It holds for the term on the last but one line in (3.24) that

$$
\begin{aligned}
&-\int_0^t \left( \int_0^s \nabla(\overline{u}_n^h - u), \nabla(u - \pi_h u) \right) \\
&\leq \int_0^t \left\| \int_0^s \nabla(\overline{u}_n^h - u) \right\|^2 + \int_0^t \|\nabla(u - \pi_h u)\|^2 .
\end{aligned}
$$

The analogous upper bounds can be shown also for the last two terms on the RHS in (3.24). We collect the estimates, apply Gronwall's theorem and the proof is complete. □

Let us consider first-order Lagrange finite elements for the space discretization (see Example A.1 ). Assume that $T_h$ is a regular family of triangulation of the domain $\Omega$ and the boundary $\Gamma$ in the sense of Definition A.21. It holds true that all the finite elements $(K, P_K, \Sigma_K), K \in \cup_h T_h$ are of class $C^0$ and for the reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ satisfies the inclusions

$$P_1(\hat{K}) \subset \hat{P} \subset H^1(\hat{K}),$$

where $P_1(\hat{K})$ stands the space first-order polynomials on $\hat{K}$. If $\dim(\hat{K}) \leq 3$, we have also the compact Sobolev embedding

$$H^2(\hat{K}) \hookrightarrow\hookrightarrow C(\hat{K}).$$

Then, it follows from [31, Theorem 3.2.1] (see Theorem A.20), that for a sufficiently regular function $u$

$$\|u - \pi_h u\|_{H^1(\Omega)} \leq Ch|u|_{H^2(\Omega)} \tag{3.25}$$

where $\pi_h$ is defined by (A.6). The same assertion can be deduced for the finite elements on the boundary $\Gamma$

$$\|u - \pi_h u\|_{H^1(\Gamma)} \leq Ch|u|_{H^2(\Gamma)}. \tag{3.26}$$

It is easy to see in light of (3.25) and (3.26) that

$$
\begin{aligned}
&\sqrt{\int_0^T \|u - \pi_h u\|^2} + \sqrt{\int_0^T \|u - \pi_h u\|_\Gamma^2} \\
&+ \int_0^T \left[ \|\nabla(u - \pi_h u)\|^2 + \|u - \pi_h u\|_\Gamma^2 + \|\nabla_\Gamma(u - \pi_h u)\|_\Gamma^2 \right] \\
&\leq C(h + h^2) \\
&\leq Ch,
\end{aligned}
$$

which proves the corollary of Theorem 3.3 stated below.

**Corollary 3.1.** *Let the assumptions of Theorem 3.3 hold. Suppose that* $d \leq 3$ *and* $u \in L^2((0,T), H^2(\Omega)) \cap L^2((0,T), H^2(\Gamma))$. *Then for any* $0 < \tau < \tau_0$ *and* $0 < h < h_0 < 1$ *we have*

$$\begin{aligned}
&\int_0^T \left\|\theta(\overline{u}_n^h) - \theta(u)\right\|^2 + \int_0^T \left\|\overline{u}_n^h - u\right\|_\Gamma^2 \\
&+ \left\|\int_0^T \nabla(\overline{u}_n^h - u)\right\|^2 + \left\|\int_0^T (\overline{u}_n^h - u)\right\|_\Gamma^2 + \left\|\int_0^T \nabla(\overline{u}_n^h - u)\right\|_\Gamma^2 \\
&\leq C\big(\tau + h\big).
\end{aligned}$$

## 3.5 Numerical experiments

The last section of this chapter is devoted to some numerical experiments to support the theoretical conclusions. We investigate the convergence rate of the numerical solution to a given exact one. We compare, in particular, the time discretization error $E$ in the sense of Theorem 2

$$E^2 = \int_0^T \|u_{exact} - u_{numerical}\|^2 \, \mathrm{d}t + \int_0^T \|u_{exact} - u_{numerical}\|_\Gamma^2 \, \mathrm{d}t$$

with respect to the decreasing time step $\tau$ and the space discretization parameter $h$.

The computational scheme follows the theoretical analysis. The backward Euler method is applied to discretize the problem in time. The space discretization is carried out by the finite element method with the first-order Lagrange finite elements. The resultant nonlinear system is solved by fixed point method. We have used FEniCS software [78].

Figure 3.1 displays the outcome for the functions $\theta(u) = u^{1/2}$, $\beta(u) = u^{1/2}$ and the exact solution $u_{exact} = (1+t^2)(2+xyz)$ on the unit ball domain $\Omega = \{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 < 1\}$.

Figure 3.2 displays the outcome for the functions $\theta(u) = u^{1/2}$, $\beta(u) = u^{1/2}$ and the exact solution $u_{exact} = (1+t^2)(2+\sin(xy))$ on the unit disc $\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ for a fine time step $\tau = 0.001$.

Inspecting the numerical results depicted in Figure 3.1, we see that the time discretization error is $\mathcal{O}(\tau)$ for $\tau > 10^{-5}$. For $\tau < 10^{-5}$ the space discretization becomes more dominant which causes the stabilization of total error.

Figure 3.2 shows the space discretization error of the order $\mathcal{O}\left(h^{1.9}\right)$ for $h > 10^{-1}$. If $h < 10^{-1}$ then the time discretization error becomes more dominant, which is why we get the flat part on the plot.

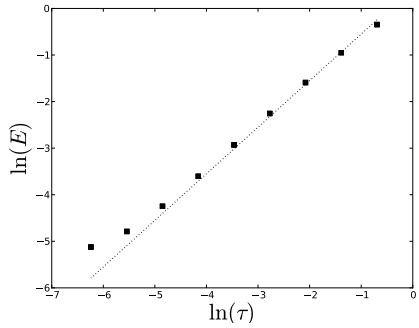



Figure 3.1: Absolute error with respect to time step $\tau$ in log-log scale. The dotted line has the slope one

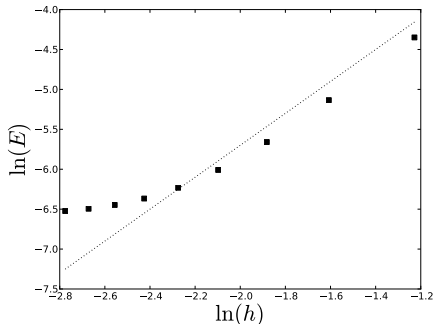



Figure 3.2: Absolute error with respect to the space discretization parameter $h$ and $\tau = 0.001$ in log-log scale. The dotted line has the slope two

# Chapter 4

# An eddy current problem with a nonlinear impedance-like boundary condition

In this chapter we turn our attention to boundary conditions in electromagnetism. We will study an eddy current electromagnetic problem with a nonlinear impedance-like boundary condition.

## 4.1 The Maxwell equations

Electromagnetic phenomena are governed by the set of the Maxwell equations

$$\begin{aligned} \nabla \times \boldsymbol{h}(x,t) &= \partial_t \boldsymbol{d}(x,t) + \boldsymbol{j}(x,t), \quad & \nabla \cdot \boldsymbol{d}(x,t) &= \rho(x,t), \\ \nabla \times \boldsymbol{e}(x,t) &= -\partial_t \boldsymbol{b}(x,t), & \nabla \cdot \boldsymbol{b}(x,t) &= 0, \end{aligned} \tag{4.1}$$

where $x = (x_1, x_2, x_3)^\top \in \mathbb{R}^3$ is the space variable and $t$ is the time variable. The Maxwell equations describe how the electric fields $\boldsymbol{e}$ and $\boldsymbol{d}$ and the magnetic fields $\boldsymbol{h}$ and $\boldsymbol{b}$ interact with each other and with electric charges $\rho$ and currents $\boldsymbol{j}$ [1]. The whole system (4.1) was first formulated by James Clark Maxwell [80]. The above vector calculus formulation with the curl [2] and divergence operator is thanks to Oliver Heaviside and Josiah Willard Gibbs.

[1] We use bold lowercase letters to denote vector electromagnetic quantities as a compromise between vector calculus and functional analysis notation. The capital letters will be primarly reserved for sets or function spaces and lowercases for their elements, for instance $\boldsymbol{h} \in \boldsymbol{V}$ or $x \in \Omega$. We sometimes write bold $\boldsymbol{x}$ to emphasize that it is a vector.

[2] In Cartesian coordinates $\nabla \times \boldsymbol{u} = \left( \frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right)$

The first equation in (4.1) is Ampère's law with Maxwell's correction, which states that the magnetic field can generated by electrical current or by changing electric fields. It was first discovered by André-Marie Ampère and then improved by J. C. Maxwell who added the displacement current term $\partial_t \boldsymbol{d}$. The equation below is called Faraday's law. It descibes how a time varying magnetic field creates an electric field. This phenomenon is known as a magnetic induction. The next equation, Gauss's law, relates the distribution of the electric charge $\rho$ to the resulting electric field. The last equation, sometimes called Gauss's law for magnetism, states that there are no "magnetic charges", analogous to electric charges.

The equations of electromagnetism are not complete without the constitutive laws for the electric and magnetic fields

$$\boldsymbol{d} = \epsilon \boldsymbol{e},$$
$$\boldsymbol{b} = \mu \boldsymbol{h}.$$

The permittivity $\epsilon$ is a measure of the ability of a material to resist the formation of an electric field within it. The permeability $\mu$ is a measure of magnetization of a material in response to a magnetic field. Another important constitutive relation is Ohm's law

$$\boldsymbol{j} = \sigma \boldsymbol{e},$$

which states that the current through a conductor is directly proportional to the electric field with the material dependent conductivity $\sigma$. The constitutive laws can bring nonlinear dependences into the (basically linear) Maxwell equations. They describes material's response to the electric and magnetic field, which is often nonlinear. We mention here the reference [46], which is devoted to the mathematical modelling of electromagnetic solids. Diffusion of electromagnetic fields in magnetically nonlinear conductors and electrically nonlinear superconductors is covered in the book [81]. Among other topics, the book discusses a power law approximation of a magnetization curve

$$\boldsymbol{b} = k\boldsymbol{h}^{1/n}, \qquad n > 1.$$

The Maxwell equations can be elegantly formulated in the exterior differential form language:

$$\mathrm{d}\mathcal{F} = 0, \quad \mathrm{d} \star \mathcal{F} = \mathcal{J},$$

where $\mathcal{F} = \mathrm{d}t \wedge \mathcal{E} - \mathcal{B}$ is the electromagnetic 2-form and $\mathcal{J}$ is the electric current 3-form in the $(3+1)$-dimensional spacetime manifold. The exterior derivative $\mathrm{d}$ and the Hodge dual operator $\star$ are standard tools of differential geometry (see [49, Chapter 16]). Let us leave the abstract spacetime and return to the standard three dimensional Euclidean space with separate time variable $t$ as in the paper [120]. The above equations read in a more familiar form as

$$\mathrm{d}\mathcal{H} = \partial_t \mathcal{D} + \mathcal{J}, \qquad \mathrm{d}\mathcal{D} = \rho,$$
$$\mathrm{d}\mathcal{E} = -\partial_t \mathcal{B}, \qquad \mathrm{d}\mathcal{B} = 0.$$

They are linked by the constitutive laws $\mathcal{D} = \epsilon \star \mathcal{E}$ and $\mathcal{B} = \mu \star \mathcal{H}$. The electromagnetic fields $\mathcal{E}$ and $\mathcal{H}$ are here 1-forms, and $\mathcal{D}, \mathcal{B}$ and $\mathcal{J}$ are 2-forms. We refer the interested reader to the book [96] for a general introduction to electromagnetism discussed in exterior differential form calculus notation.

Treating electromagnetic fields as differential forms and not as vector fields has many advantages. It offers a superior insight into the geometrical nature of electromagnetism. The differential form approach provides a natural interpretation of edge finite elements. An in-depth exposition can found in the book [18] by A. Bossavit who has, among others, pioneered this approach. The framework of discrete differential forms have proven since then to be very practical and useful in computational electromagnetism [63].

## 4.2 Boundary conditions in electromagnetism

The Maxwell equations have to be accompanied by adequate boundary conditions on the boundary of the study region $\Omega$. We will briefly go over a few of them.

We begin with boundary conditions at interfaces between different media. They follow straight from the Maxwell equations. We draw their derivation from the classical reference [67]. Let $V$ be a finite volume in space with the boundary $S$ and denote by $\boldsymbol{n}$

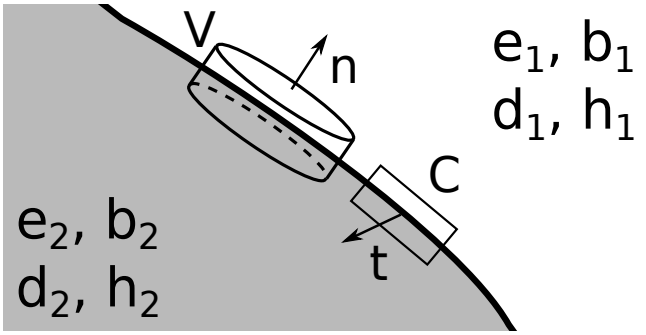



Figure 4.1: Schematic diagram of boundary surface (heavy line) between different media

the normal unit vector pointing outward from the enclosed volume. Then the divergence theorem applied to the first two Maxwell equations in (4.1) yields the integral statements

$$\oint_S \boldsymbol{d} \cdot \boldsymbol{n} \, \mathrm{d}s = \int_V \rho \, \mathrm{d}x \quad \text{and} \quad \oint_S \boldsymbol{b} \cdot \boldsymbol{n} \, \mathrm{d}s = 0. \tag{4.2}$$

Similarly, let $C$ be a closed contour in the space and $S$ an open surface. Applying the Stokes theorem to the last two Maxwell equations in (4.1) gives

$$\oint_C \boldsymbol{h} \cdot \, \mathrm{d}\boldsymbol{l} = \int_{S'} (\partial_t \boldsymbol{d} + \boldsymbol{j}) \cdot \boldsymbol{n} \, \mathrm{d}s \quad \text{and} \quad \oint_C \boldsymbol{e} \cdot \, \mathrm{d}\boldsymbol{l} = \int_{S'} -\partial_t \boldsymbol{b} \cdot \boldsymbol{n} \, \mathrm{d}s. \tag{4.3}$$

Consider now the geometrical arrangement shown in Figure 4.1. An infinitesimal pillbox $V$ straddles the boundary surface between two media with different electromagnetic properties. Similarly, the infinitesimal contour $C$ has its long arms on either side of the boundary and is oriented so that the normal to its spanning surface is tangent to the interface. We first apply the integral statements (4.2) to the volume of the pillbox. In the limit of a very shallow pillbox having the width $\delta s$ , the side surface does not contribute to the integrals on the left of (4.2). Only top and bottom contribute. If they are parallel, tangent to the surface, we obtain

$$\oint_S \boldsymbol{d} \cdot \boldsymbol{n} \, \mathrm{d}s = (\boldsymbol{d}_2 - \boldsymbol{d}_1) \cdot \boldsymbol{n} \, \delta s \quad \text{and} \quad \int_V \rho \, \mathrm{d}x = \sigma \, \delta s,$$

where $\sigma$ is the surface charge density. Thus the normal components of $\boldsymbol{d}$ and $\mathbf{b}$ on either side of the boundary surface are related according to

$$(\boldsymbol{d}_2 - \boldsymbol{d}_1) \cdot \boldsymbol{n} = \sigma \quad \text{and} \quad (\boldsymbol{b}_2 - \boldsymbol{b}_1) \cdot \boldsymbol{n} = 0.$$

In an analogous manner the infinitesimal Stokesian loop can be used to determine the discontinuities of the tangential components of $\boldsymbol{e}$ and $\boldsymbol{h}$. We obtain from the second formula in (4.3)

$$\oint_C \boldsymbol{h} \cdot \, \mathrm{d}\boldsymbol{l} = (\boldsymbol{t} \times \boldsymbol{n}) \cdot (\boldsymbol{e}_2 - \boldsymbol{e}_1) \, \delta l \quad \text{and} \quad \int_{S'} (\partial_t \boldsymbol{d} + \boldsymbol{j}) \cdot \boldsymbol{n} \, \mathrm{d}s = \boldsymbol{k} \cdot \boldsymbol{t} \, \delta l$$

by neglecting the contributions of short arms of $C$. The symbol $\boldsymbol{k}$ stands for the surface current density. The tangential components of $\boldsymbol{e}$ and $\boldsymbol{h}$ on either side of the boundary are therefore related by

$$(\boldsymbol{e}_2 - \boldsymbol{e}_1) \times \boldsymbol{n} = \boldsymbol{k} \quad \text{and} \quad (\boldsymbol{h}_2 - \boldsymbol{h}_1) \times \boldsymbol{n} = \mathbf{0}.$$

A classical example of the boundary condition of this type is the perfect electric conductor boundary condition, which enforces the tangential electric field or normal magnetic flux to be equal to zero at the boundary of the considered domain

$$\boldsymbol{n} \times \boldsymbol{e} = \mathbf{0} \quad \text{or} \quad \boldsymbol{n} \cdot \boldsymbol{b} = 0.$$

It says that the electromagnetic fields do not penetrate the conductor on the other side of the boundary. The so-called skin depth is zero. The associated tendency of electromagnetic fields to concentrate near the surface is known as the skin effect. In practice, real materials have finite conductivity. They allow the diffusion of electromagnetic fields into themselves (the skin depth is no more zero) which has to be sometimes taken into account.

The impedance boundary condition is an approximate boundary condition to capture the skin effect in the thin layer. The classical impedance boundary condition is derived

from the full time-harmonic Maxwell equations. It is obtained as a certain ratio between electric and magentic field component of their particular solution. The book [123] represents a comprehensive survey on surface impedance boundary conditions. The most of the literature discusses the impedance boundary condition in the time-harmonic regime, in the so-called frequency domain, [99]. One can use the inverse Laplace or Fourier transform to get back to the time domain as in [65]. The papers [59] and [38] deal with generalized impedance BCs directly in time domain. They model a thin coating for electromagnetic scattering problems. To derive the so-called effective boundary condition, they asymptoticaly expand the solution with respect to the thickness of the thin layer and use the decomposition of the curl operator on the thin layer with respect to normal $\boldsymbol{n}$. We note that this becomes in planar case

$$\begin{aligned} \nabla \times \boldsymbol{\varphi} &= \overrightarrow{\mathrm{rot}}_\Gamma(\boldsymbol{\varphi} \cdot \boldsymbol{n}) + (\mathrm{rot}_\Gamma \boldsymbol{\varphi})\boldsymbol{n} - \partial_\nu(\boldsymbol{\varphi} \times \boldsymbol{n}) \\ &\equiv (\nabla_\Gamma(\boldsymbol{\varphi} \cdot \boldsymbol{n})) \times \boldsymbol{n} + \mathrm{div}_\Gamma(\boldsymbol{\varphi} \times \boldsymbol{n})\boldsymbol{n} - \partial_\nu(\boldsymbol{\varphi} \times \boldsymbol{n}), \end{aligned} \tag{4.4}$$

where $\nabla_\Gamma$ is the surface gradient, $\mathrm{div}_\Gamma$ is the surface divergence [3], and $\nu$ is the coordinate in the direction $\boldsymbol{n}$.

Impedance-type interface conditions has also been studied in the context of thin shells [76]. The authors of [57] has proposed a time-domain extension of the frequency-domain thin-shell approach. A instructive derivation of impedance boundary conditions for thin shell in the linear case can be found in [53, Appendix B]. Let us follow this approach to derive an impedance-type interface condition in a nonlinear case and without the knowledge of the exact solution. Consider the situation on Figure 4.2, where the thin layer $\Omega_t$ is on the top of the domain $\Omega$. Suppose that the normal component of the

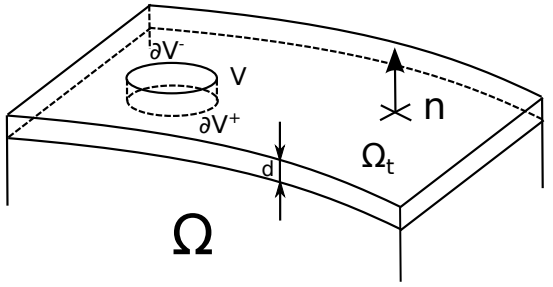



Figure 4.2: Thin shell on the top of the domain $\Omega$

electric field $\boldsymbol{e}$ can be neglected and there is no electric source in $\Omega_t$. Therefore, we set

$$\boldsymbol{e} \cdot \boldsymbol{n} = 0 \quad \text{and} \quad \mathrm{div}_\Gamma(\boldsymbol{e} \times \boldsymbol{n}) = 0 \quad \text{in } \Omega_t. \tag{4.5}$$

[3] see Chapter 2 for the definitions

The latter formula is justified by assuming the constant permittivity $\epsilon$ in the constitutive law $\boldsymbol{d} = \epsilon \boldsymbol{e}$. We now integrate Faraday's law over the test volume $V$. Together with the formula (4.4) and the assumptions (4.5), it yields

$$-\int_V \partial_t \boldsymbol{b}\,\mathrm{d}x = \int_V \nabla \times \boldsymbol{e}\,\mathrm{d}x = \int_{\partial V^+} \boldsymbol{n} \times \boldsymbol{e}\,\mathrm{d}s - \int_{\partial V^+} \boldsymbol{n} \times \boldsymbol{e}\,\mathrm{d}s.$$

The trapezoidal rule in the $\nu$-coordinate to approximate the integral containing $\partial_t \boldsymbol{b}$ then leads to

$$\boldsymbol{n} \times \boldsymbol{e}|_{\partial V^+} - \boldsymbol{n} \times \boldsymbol{e}|_{\partial V^-} = -\frac{d}{2}\left(\partial_t \boldsymbol{b}|_{\partial V^+} + \partial_t \boldsymbol{b}|_{\partial V^-}\right).$$

In the linear case the above formula coincides with the formula B.15 in [53, Appendix B] given that $d/2 = \beta$.

A related class of approximate boundary conditions is absorbing boundary conditions for electromagnetic wave problems. They are used to truncate an unbounded computational domain so that an outcoming wave is not reflected back to the domain of interest [43]. There have been proposed different families of absorbing BC, see for instance [58]. We remark that there exists a different approach to tackle this issue which uses the so-called perfectly matched layers [16].

In this chapter we study the eddy current approximation of the Maxwell equations

$$\begin{aligned} \nabla \times \boldsymbol{h} &= \boldsymbol{j} \\ \nabla \times \boldsymbol{e} &= -\partial_t \boldsymbol{b} \end{aligned} \qquad \text{in } \Omega \times (0,T), \tag{4.6}$$

with the impedance-like boundary condition

$$\boldsymbol{n} \times \boldsymbol{e} = \boldsymbol{n} \times (\partial_t \boldsymbol{b}(\boldsymbol{h}) \times \boldsymbol{n}) \qquad \text{on } \Gamma \times (0,T). \tag{4.7}$$

The boundary $\Gamma = \partial\Omega$ of the bounded domain $\Omega \subset \mathbb{R}^3$ is Lipschitz continuous and $\boldsymbol{n}$ stands for the unit outward normal vector on $\Gamma$. We assume the power law nonlinearity on the boundary

$$\boldsymbol{b}(\boldsymbol{h}) = \boldsymbol{a}(\boldsymbol{h}) = a(|\boldsymbol{h}|)\boldsymbol{h} = |\boldsymbol{h}|^{\alpha-1}\boldsymbol{h}, \quad \alpha \in (0,1]. \tag{4.8}$$

The boundary condition (4.7) is a dissipative boundary condition between the tangential components of $\boldsymbol{h}$ and $\boldsymbol{e}$, which corresponds to a non-perfect contact at the boundary. It means that the material on one side of the boundary does not allow the field to penetrate without loosing the energy. Let us take a closer look at the normal component of the Poynting vector $\boldsymbol{e} \times \boldsymbol{h}$ on the boundary to see it. It is easy to check that $\partial_t |\boldsymbol{h}|^2 = 2\boldsymbol{h} \cdot \partial_t \boldsymbol{h}$, and

$$\partial_t \left(|\boldsymbol{h}|^\beta\right) = \partial_t \left(\left(|\boldsymbol{h}|^2\right)^{\frac{\beta}{2}}\right) = \beta |\boldsymbol{h}|^{\beta-2} \boldsymbol{h} \cdot \partial_t \boldsymbol{h}.$$

It follows from the preceding that

$$\partial_t \left(|\boldsymbol{h}|^\beta \boldsymbol{h}\right) = \beta |\boldsymbol{h}|^{\beta-2} (\boldsymbol{h} \cdot \partial_t \boldsymbol{h}) \boldsymbol{h} + |\boldsymbol{h}|^\beta \partial_t \boldsymbol{h}$$

and

$$\partial_t \left(|\boldsymbol{h}|^\beta \boldsymbol{h}\right) \cdot \boldsymbol{h} = \frac{1+\beta}{2+\beta} \partial_t \left(|\boldsymbol{h}|^{\beta+2}\right).$$

Integration in time yields

$$\begin{aligned}
\int_0^s \boldsymbol{e} \times \boldsymbol{h} \cdot \boldsymbol{n} \, \mathrm{d}t &= \int_0^s \boldsymbol{n} \times \boldsymbol{e} \cdot \boldsymbol{h} \, \mathrm{d}t \\
&= \int_0^s \partial_t \left[ |\boldsymbol{h} \times \boldsymbol{n}|^{\alpha-1} \boldsymbol{h} \times \boldsymbol{n} \right] \cdot (\boldsymbol{h} \times \boldsymbol{n}) \, \mathrm{d}t \\
&= \frac{\alpha}{1+\alpha} \int_0^s \partial_t \left( |\boldsymbol{h} \times \boldsymbol{n}|^{1+\alpha} \right) \mathrm{d}t \\
&= \frac{\alpha}{1+\alpha} \left( |\boldsymbol{h}(s) \times \boldsymbol{n}|^{1+\alpha} - |\boldsymbol{h}(0) \times \boldsymbol{n}|^{1+\alpha} \right),
\end{aligned}$$

which is valid at each point of the boundary $\Gamma$. We see that if $|\boldsymbol{h}(0) \times \boldsymbol{n}| = 0$, then (4.7) is locally absorbing in the sense of the definition [46, §7.12]. That is

$$\int_0^s \boldsymbol{e}(\boldsymbol{x}, t) \times \boldsymbol{h}(\boldsymbol{x}, t) \cdot \boldsymbol{n} \, \mathrm{d}t \geq 0, \qquad \boldsymbol{x} \in \Gamma.$$

Nonlinear boundary conditions for Maxwell equations have been studied e.g. in [40, 41, 103]. The BC there do not contain the time derivative. For a full Maxwell system with an evolution BC we refer to [127], where the error of a time-discretization has been studied.

We assume a linear relation between the magnetic fields $\boldsymbol{b}$ and $\boldsymbol{h}$ in the domain $\Omega$ and the linear Ohm law for the current density $\boldsymbol{j}$

$$\boldsymbol{b} = \mu \boldsymbol{h}, \qquad \boldsymbol{j} = \sigma \boldsymbol{e}.$$

We use the scaling $\mu = \sigma = 1$ for ease of explanation.

Eliminating the electric field $\boldsymbol{e}$, we are left with the following nonlinear parabolic problem

$$\begin{aligned}
\partial_t \boldsymbol{h} + \nabla \times \nabla \times \boldsymbol{h} &= \boldsymbol{0}, && \text{in } \Omega \times (0,T), \\
\boldsymbol{n} \times (\nabla \times \boldsymbol{h}) &= \boldsymbol{n} \times \partial_t \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}), && \text{on } \Gamma \times (0,T), \\
\boldsymbol{h}(\boldsymbol{x}, 0) &= \boldsymbol{h}_0 && \text{in } \Omega.
\end{aligned} \tag{4.9}$$

The main goal of this chapter is to demonstrate the unique solvability of (4.9) in appropriate spaces and establish error estimates of time and space discretization.

## 4.3 Weak formulation

In this section we formulate the problem (4.9) in the weak sense. We first extend the relevant notation to vector-valued functions. We write $\boldsymbol{L}^2(\Omega)$ for the vector space $(L^2(\Omega))^3$. Let $\boldsymbol{u} = (u_1, u_2, u_3)^\top \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{v} = (v_1, v_2, v_3)^\top \in \boldsymbol{L}^2(\Omega)$. The notation for its inner product and the associated norm is

$$(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \boldsymbol{u}(x) \cdot \boldsymbol{v}(x) \,\mathrm{d}x = \sum_{j=1}^{3} \int_\Omega u_j(x) v_j(x) \,\mathrm{d}x \quad \text{and} \quad \|\boldsymbol{u}\| = \sqrt{(\boldsymbol{u}, \boldsymbol{u})}.$$

The notation for the boundary space $\boldsymbol{L}^2(\Gamma) = (L^2(\Gamma))^3$ is obvious

$$(\boldsymbol{u}, \boldsymbol{v})_\Gamma = \int_\Gamma \boldsymbol{u}(S) \cdot \boldsymbol{v}(S) \,\mathrm{d}S = \sum_{j=1}^{3} \int_\Gamma u_j(S) v_j(S) \,\mathrm{d}S \quad \text{and} \quad \|\boldsymbol{u}\|_\Gamma = \sqrt{(\boldsymbol{u}, \boldsymbol{u})_\Gamma},$$

where $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{L}^2(\Gamma)$.

We introduce the standard space of three-dimensional vector functions with their curl in $\boldsymbol{L}^2$-space

$$\boldsymbol{H}(\mathrm{curl}; \Omega) = \{\boldsymbol{\varphi} \in \mathbf{L}^2(\Omega) : \quad \nabla \times \boldsymbol{\varphi} \in \mathbf{L}^2(\Omega)\}.$$

It is equipped with the norm

$$\|\boldsymbol{\varphi}\|_{\boldsymbol{H}(\mathrm{curl};\Omega)} = \sqrt{\|\boldsymbol{\varphi}\|^2 + \|\nabla \times \boldsymbol{\varphi}\|^2}.$$

This Hilbert space is a natural $L^2$-setting for many boundary value problems in electromagnetism. We omit a detailed description of the traces for $\boldsymbol{H}(\mathrm{curl}; \Omega)$ by referring to the paper [22]. More on functional analysis framework for electromagnetism can be found in the book [27]. We define the space $\boldsymbol{V}$

$$\boldsymbol{V} = \{\boldsymbol{\varphi} \in \boldsymbol{H}(\mathrm{curl}; \Omega) : \quad \boldsymbol{\varphi} \times \boldsymbol{n} \in \boldsymbol{L}^{1+\alpha}(\Gamma)\} \tag{4.10}$$

to involve the power law (4.8) in the boundary condition (4.7). The space $\boldsymbol{L}^{1+\alpha}(\Gamma)$ consists of vector-valued Lebesgue measurable functions on $\Gamma$ with the norm

$$\|\boldsymbol{\varphi}\|_{\boldsymbol{L}^{1+\alpha}(\Gamma)} = \left( \int_\Gamma |\boldsymbol{\varphi}(x)|^{1+\alpha} \,\mathrm{d}S \right)^{1/(1+\alpha)}.$$

The definition (4.10) enables us to avoid working with usual traces spaces of $\boldsymbol{H}(\mathrm{curl}; \Omega)$. The space $\boldsymbol{V}$ is a Banach space endowed with the graph norm

$$\|\boldsymbol{\varphi}\|_{\boldsymbol{V}} = \sqrt{\|\boldsymbol{\varphi}\|^2_{\boldsymbol{H}(\mathrm{curl};\Omega)} + \|\boldsymbol{\varphi} \times \boldsymbol{n}\|^2_{\boldsymbol{L}^{1+\alpha}(\Gamma)}}.$$

A weak solution of the boundary value problem (4.9) has to satisfy

$$(\partial_t \boldsymbol{h}, \boldsymbol{\varphi}) + (\nabla \times \boldsymbol{h}, \nabla \times \boldsymbol{\varphi}) + (\partial_t \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma = 0, \qquad \boldsymbol{h}(0) = \boldsymbol{h}_0 \tag{4.11}$$

for any function $\boldsymbol{\varphi}$ from the space $\boldsymbol{V}$ and for almost every time $t$ from the open interval $(0, T)$.

The first theorem ensures uniqueness of a solution to the problem (4.11)

**Theorem 4.1.** *For any initial data $\boldsymbol{h}_0 \in \boldsymbol{V}$ there exists at most one solution $\boldsymbol{h}$ of the problem (4.11) satisfying the conditions: $\boldsymbol{h} \in L_2((0,T), \boldsymbol{V})$, $\partial_t \boldsymbol{h} \in L_2((0,T), \boldsymbol{L}_2(\Omega))$ and $\partial_t \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}) \in L_2((0,T), \boldsymbol{V}^*)$.*

*Proof.* We begin by showing the crucial characteristic of the problem and that is the monotonicity of the vector field $\boldsymbol{a} : \mathbb{R}^3 \to \mathbb{R}^3$ which is defined by the power law (4.8). The directional derivative of $\boldsymbol{a}$ in a direction $\boldsymbol{u} \in \mathbb{R}^3$ can be expanded by the product rule

$$\langle \operatorname{grad} \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{u} \rangle = \langle \operatorname{grad} a(|\boldsymbol{x}|)\boldsymbol{x}, \boldsymbol{h} \rangle = a'(|\boldsymbol{x}|)\frac{\boldsymbol{x} \cdot \boldsymbol{u}}{|\boldsymbol{x}|}\boldsymbol{x} + a(|\boldsymbol{x}|)\boldsymbol{u}.$$

Let $\theta \in (0,1)$. We find using the mean value theorem and the above identity that

$$\begin{aligned}
[\boldsymbol{a}(\boldsymbol{x}+\boldsymbol{u}) - \boldsymbol{a}(\boldsymbol{x})] \cdot \boldsymbol{u} &= \langle \operatorname{grad} \boldsymbol{a}(\boldsymbol{x} + \theta\boldsymbol{u}), \boldsymbol{u} \rangle \cdot \boldsymbol{u} \\
&= a(|\boldsymbol{x} + \theta\boldsymbol{u}|)|\boldsymbol{u}|^2 + a'(|\boldsymbol{x} + \theta\boldsymbol{u}|)\frac{((\boldsymbol{x} + \theta\boldsymbol{u}) \cdot \boldsymbol{u})^2}{|\boldsymbol{x} + \theta\boldsymbol{u}|} \\
&\geq \Big(a(|\boldsymbol{x} + \theta\boldsymbol{u}|) - |a'(|\boldsymbol{x} + \theta\boldsymbol{u}|)||\boldsymbol{x} + \theta\boldsymbol{u}|\Big)|\boldsymbol{u}|^2 \\
&\geq \alpha|\boldsymbol{x} + \theta\boldsymbol{u}|^{\alpha-1}|\boldsymbol{u}|^2 \\
&\geq 0,
\end{aligned}$$

which proves the monotonicity of $\boldsymbol{a}$.

We now proceed to the uniqueness of a solution. Let us suppose, to the contrary, that there are two different solutions $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ of the problem (4.11). We subtract the weak formulation for $\boldsymbol{h}_2$ from the one for $\boldsymbol{h}_1$ and integrate in time. The subsequent substitution $\boldsymbol{\varphi} = \boldsymbol{h}_1 - \boldsymbol{h}_2$ and integration over the interval $(0, T)$ yields

$$\begin{aligned}
&\int_0^T \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|^2 \,\mathrm{d}t \\
&+ \int_0^T \left( \int_0^t \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2) \,\mathrm{d}s, \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)(t) \right) \mathrm{d}t \\
&+ \int_0^T (\boldsymbol{a}(\boldsymbol{h}_1 \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}_2 \times \boldsymbol{n}), \boldsymbol{h}_1 \times \boldsymbol{n} - \boldsymbol{h}_2 \times \boldsymbol{n})_\Gamma \,\mathrm{d}t = 0.
\end{aligned}$$

Integration by parts shows that

$$\int_0^T \left( \int_0^t \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)\,\mathrm{d}s, \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)(t) \right) \mathrm{d}t$$
$$= \frac{1}{2} \left\| \int_0^T \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)\,\mathrm{d}t \right\|^2 .$$

We conclude by monotonicity of $\boldsymbol{a}$ that the both functions $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are identical

$$\int_0^T \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|^2\,\mathrm{d}t + \frac{1}{2} \left\| \int_0^T \nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)\,\mathrm{d}t \right\|^2 \leq 0,$$

which is the contradiction. □

In order to show existence of the solution to the problem (4.11), we employ Rothe's method. We divide the time interval $[0, T]$ into $n$ equidistant subintervals $[t_{i-1}, t_i]$ for $t_i = i\tau$, where $\tau = T/n$. As always,

$$\boldsymbol{u}_i = \boldsymbol{u}(t_i), \qquad \delta \boldsymbol{u}_i = \frac{\boldsymbol{u}_i - \boldsymbol{u}_{i-1}}{\tau}.$$

A time-discretized weak formulation of (4.11) reads as

$$(\delta \boldsymbol{h}_i, \boldsymbol{\varphi}) + (\nabla \times \boldsymbol{h}_i, \nabla \times \boldsymbol{\varphi}) + (\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma = 0, \qquad \boldsymbol{\varphi} \in \boldsymbol{V}, \qquad (4.12)$$
$$\boldsymbol{h}_0 = \boldsymbol{h}_0$$

for every $i = 1, \ldots, n$. The next lemma guarantees the existence of a weak solution on every time step.

**Lemma 4.1.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{V}$, then there exist a uniquely determined $\boldsymbol{h}_i \in \boldsymbol{V}$ solving (4.12) for any index $i = 1, \ldots, n$.*

*Proof.* The proof is a simple application of monotone operator theory. With the formula (4.12) in mind, we first define the nonlinear mapping $A : \boldsymbol{V} \to \boldsymbol{V}^*$

$$\langle A(\boldsymbol{h}), \boldsymbol{\varphi} \rangle = \left( \frac{\boldsymbol{h}}{\tau}, \boldsymbol{\varphi} \right) + (\nabla \times \boldsymbol{h}, \nabla \times \boldsymbol{\varphi}) + \left( \frac{\boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n})}{\tau}, \boldsymbol{\varphi} \times \nu \right)_\Gamma,$$

where $\boldsymbol{\varphi} \in \boldsymbol{V}$. It can be checked that this mapping is hemicontinuous. It is also easy to see that

$$\langle A(\boldsymbol{h}), \boldsymbol{h} \rangle = \frac{\|\boldsymbol{h}\|^2}{\tau} + \|\nabla \times \boldsymbol{h}\|^2 + \int_\Gamma \frac{|\boldsymbol{h} \times \boldsymbol{n}|^{\alpha+1}}{\tau}\,\mathrm{d}S$$
$$\geq \frac{\|\boldsymbol{h}\|^2_{\boldsymbol{H}(\mathrm{curl};\Omega)} + \|\boldsymbol{h} \times \boldsymbol{n}\|^{1+\alpha}_{\boldsymbol{L}^{1+\alpha}(\Gamma)}}{\tau}$$

for $0 < \tau < 1$, which means that the mapping $A$ is coercive, i.e

$$\frac{\langle A(\boldsymbol{h}), \boldsymbol{h}\rangle}{\|\boldsymbol{h}\|_{\boldsymbol{V}}} \to \infty \quad \text{for } \|\boldsymbol{h}\|_{\boldsymbol{V}} \to \infty.$$

The strict monotonicity of the mapping $A$ follows from the monotonicity of the vector field $\boldsymbol{a}$

$$\begin{aligned}\langle A(\boldsymbol{h}_1) - A(\boldsymbol{h}_2), \boldsymbol{h}_1 - \boldsymbol{h}_2\rangle =& \frac{\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|^2}{\tau} + \|\nabla \times (\boldsymbol{h}_1 - \boldsymbol{h}_2)\|^2 \\ &+ (\boldsymbol{a}(\boldsymbol{h}_1 \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}_2 \times \boldsymbol{n}), (\boldsymbol{h}_1 - \boldsymbol{h}_2) \times \boldsymbol{n})_\Gamma \qquad (4.13)\\ \geq& \frac{\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|^2_{\boldsymbol{H}(\mathrm{curl};\Omega)}}{\tau}.\end{aligned}$$

The theory of monotone operators (see Theorem A.10) then implies that for any $i = 1, \ldots, n$ there exists the unique $\boldsymbol{h}_i \in \boldsymbol{V}$ such that

$$\langle A(\boldsymbol{h}_i), \boldsymbol{\varphi}\rangle = \frac{1}{\tau}\left[(\boldsymbol{h}_{i-1}, \boldsymbol{\varphi}) + (\boldsymbol{a}(\boldsymbol{h}_{i-1} \times \boldsymbol{n}), \boldsymbol{\varphi})_\Gamma\right], \qquad \forall \boldsymbol{\varphi} \in \boldsymbol{V}.$$

□

## 4.4 A priori estimates

This section contains auxilliary results. They concern mainly a priori estimates for the solution of discretized problem (4.12) which will help us later. We first formulate a technical lemma ([103, Lemma 2.3]).

**Lemma 4.2.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a nonnegative continuous function such that $G(s) := g(s)s$ is monotonically increasing. Let $\Phi_G$ be a primitive function of $G$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3$ we have*

$$\Phi_G(|\boldsymbol{y}|) - \Phi_G(|\boldsymbol{x}|) \leq g(|\boldsymbol{y}|)\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x}).$$

*Proof.* It follows straight from the mean value theorem and Cauchy inequality that

$$\begin{aligned}\Phi_G(|\boldsymbol{y}|) - \Phi_G(|\boldsymbol{x}|) &= \int_{|\boldsymbol{x}|}^{|\boldsymbol{y}|} g(s)s\,\mathrm{d}s = g(\theta)\theta\big(|\boldsymbol{y}| - |\boldsymbol{x}|\big)\\ &\leq g(|\boldsymbol{y}|)|\boldsymbol{y}|\big(|\boldsymbol{y}| - |\boldsymbol{x}|\big) = g(|\boldsymbol{y}|)|\boldsymbol{y}|\big(|\boldsymbol{y}|^2 - |\boldsymbol{y}||\boldsymbol{x}|\big)\\ &\leq g(|\boldsymbol{y}|)\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x}),\end{aligned}$$

where $\theta$ lies between $|\boldsymbol{x}|$ and $|\boldsymbol{y}|$. □

The two next lemmas give us a priori information about $\boldsymbol{h}_i$ and $\delta\boldsymbol{h}_i$. Since we do not suppose the strict monotonicity of $\boldsymbol{a}$, we can not get any information about $\delta\boldsymbol{h}_i$ as it will turn out from the proofs.

**Lemma 4.3.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{L}_2(\Omega), \boldsymbol{h}_0 \times \boldsymbol{n} \in \boldsymbol{L}_{1+\alpha}(\Gamma)$ and $\boldsymbol{h}_i$ is the solution of (4.12). Then there exists a positive constant $C$ such that*

$$\|\boldsymbol{h}_j\|^2 + \sum_{i=1}^{j} \|\boldsymbol{h}_i - \boldsymbol{h}_{i-1}\|^2 + \sum_{i=1}^{j} \|\nabla \times \boldsymbol{h}_i\|^2 \tau + \|\boldsymbol{h}_j \times \boldsymbol{n}\|^{1+\alpha}_{\boldsymbol{L}^{1+\alpha}(\Gamma)} \leq C$$

*for any $j = 1, \ldots, n$.*

*Proof.* Taking $\boldsymbol{\varphi} = \tau \boldsymbol{h}_i$ as a test function in (4.12) and summing it over $i = 1, \ldots, j$, we obtain

$$\sum_{i=1}^{j} (\delta\boldsymbol{h}_i, \boldsymbol{h}_i)\tau + \sum_{i=1}^{j} \|\nabla \times \boldsymbol{h}_i\|^2 \tau + \sum_{i=1}^{j} (\delta\boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \boldsymbol{h}_i \times \boldsymbol{n})_\Gamma \tau = 0\,.$$

We rewrite the first term by the Abel summation

$$\sum_{i=1}^{j} (\delta\boldsymbol{h}_i, \boldsymbol{h}_i)\tau = \frac{1}{2}\Big( \|\boldsymbol{h}_j\|^2 - \|\boldsymbol{h}_0\|^2 + \sum_{i=1}^{j} \|\boldsymbol{h}_i - \boldsymbol{h}_{i-1}\|^2 \Big)\,.$$

Lemma (4.2) with $g(s) = s^{(1-\alpha)/\alpha}$ yields the estimate for the boundary sum

$$\begin{aligned}
&\sum_{i=1}^{j} (\boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}_{i-1} \times \boldsymbol{n}), \boldsymbol{h}_i \times \boldsymbol{n})_\Gamma \\
&\quad \geq \frac{\alpha}{\alpha+1} \sum_{i=1}^{j} \int_\Gamma \left[ |\boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})|^{(\alpha+1)/\alpha} - |\boldsymbol{a}(\boldsymbol{h}_{i-1} \times \boldsymbol{n}|^{(\alpha+1)/\alpha} \right] \mathrm{d}S \\
&\quad = \frac{\alpha}{\alpha+1} \big( \|\boldsymbol{h}_j \times \boldsymbol{n}\|^{\alpha+1}_{\mathbf{L}_{1+\alpha}(\Gamma)} - \|\boldsymbol{h}_0 \times \boldsymbol{n}\|^{\alpha+1}_{\mathbf{L}_{1+\alpha}(\Gamma)} \big).
\end{aligned}$$

The proof is finished by rearranging the terms. □

**Lemma 4.4.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{V}$ and $\boldsymbol{h}_i$ is the solution of (4.12). Then there exists a positive constant $C$ such that*

$$\sum_{i=1}^{j} \|\delta\boldsymbol{h}_i\|^2 \tau + \|\nabla \times \boldsymbol{h}_j\|^2 + \sum_{i=1}^{j} \|\nabla \times (\boldsymbol{h}_i - \boldsymbol{h}_{i-1})\|^2 \leq C$$

*for any $j = 1, \ldots, n$.*

*Proof.* Setting $\boldsymbol{\varphi} = \delta \boldsymbol{h}_i \tau$ in (4.12) and adding it up for $i = 1, \dots, j$, we get

$$\sum_{i=1}^{j} \|\delta \boldsymbol{h}_i\|^2 \tau + \sum_{i=1}^{j} (\nabla \times \boldsymbol{h}_i, \nabla \times [\boldsymbol{h}_i - \boldsymbol{h}_{i-1}]) + \sum_{i=1}^{j} (\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \delta(\boldsymbol{h}_i \times \boldsymbol{n}))_\Gamma \tau = 0\,.$$

The boundary sum is nonnegative because the vector field $\boldsymbol{a}$ is monotone

$$\sum_{i=1}^{j} (\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \delta(\boldsymbol{h}_i \times \boldsymbol{n}))_\Gamma \tau \geq 0.$$

The assertion of the lemma follows immediately after rewriting the second sum by the Abel summation rule. □

The lemma below provides a missing a priori estimate for the nonlinear term $\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})$. Note that the estimate can be only established using the dual norm $\boldsymbol{V}^*$.

**Lemma 4.5.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{V}$ and $\boldsymbol{h}_i$ is the solution of (4.12). Then there exists a positive constant $C$ such that*

$$\sum_{i=1}^{j} \|\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})\|^2_{\boldsymbol{V}^*} \tau \leq C$$

*for any $j = 1, \dots, n$.*

*Proof.* The identity

$$(\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma = -(\delta \boldsymbol{h}_i, \boldsymbol{\varphi}) - (\nabla \times \boldsymbol{h}_i, \nabla \times \boldsymbol{\varphi})$$

implicitly defines a linear functional on the space $\boldsymbol{V}$. For its norm it holds true that

$$\begin{aligned}
\|(\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})\|_{\boldsymbol{V}^*} &= \sup_{\boldsymbol{\varphi} \in \boldsymbol{V}} \frac{|(\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma|}{\|\boldsymbol{\varphi}\|_{\boldsymbol{V}}} \\
&= \sup_{\boldsymbol{\varphi} \in \boldsymbol{V}} \frac{|-(\delta \boldsymbol{h}_i, \boldsymbol{\varphi}) - (\nabla \times \boldsymbol{h}_i, \nabla \times \boldsymbol{\varphi})|}{\|\boldsymbol{\varphi}\|_{\boldsymbol{V}}} \\
&\leq \sup_{\boldsymbol{\varphi} \in \boldsymbol{V}} \frac{\|\delta \boldsymbol{h}_i\| \, \|\boldsymbol{\varphi}\| + \|\nabla \times \boldsymbol{h}_i\| \, \|\nabla \times \boldsymbol{\varphi}\|}{\|\boldsymbol{\varphi}\|_{\boldsymbol{V}}} \\
&\leq \sup_{\boldsymbol{\varphi} \in \boldsymbol{V}} \frac{\big( \|\delta \boldsymbol{h}_i\| + \|\nabla \times \boldsymbol{h}_i\| \big) \, \|\boldsymbol{\varphi}\|_{\boldsymbol{H}(\mathrm{curl};\Omega)}}{\|\boldsymbol{\varphi}\|_{\boldsymbol{V}}} \\
&\leq \|\delta \boldsymbol{h}_i\| + \|\nabla \times \boldsymbol{h}_i\|\,.
\end{aligned}$$

We multiply the above inequality by $\tau$ and add it up for $i = 1, \dots, j$ to see by Lemma 4.4 that

$$\sum_{i=1}^{j} \|\delta \boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})\|^2_{\boldsymbol{V}^*} \tau = \sum_{i=1}^{j} \big( \|\delta \boldsymbol{h}_i\| + \|\nabla \times \boldsymbol{h}_i\| \big)^2 \tau \leq C,$$

which was to be proved. □

The last lemma of this section is a technical result. It claims that the vector function $\boldsymbol{a}$ is Hölder continuous. We will make use of it when dealing with full discretization of (4.11).

**Lemma 4.6.** *Assume* $0 < \alpha < 1,\ d \in \mathbb{N}$. *Then there exists a positive constant* $C$ *such that*

$$|\boldsymbol{a}(\boldsymbol{x}) - \boldsymbol{a}(\boldsymbol{y})| = \left||\boldsymbol{x}|^{\alpha-1}\boldsymbol{x} - |\boldsymbol{y}|^{\alpha-1}\boldsymbol{y}\right| \leq C|\boldsymbol{x}-\boldsymbol{y}|^{\alpha}, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \quad (4.14)$$

*Proof.* If $\boldsymbol{x} = \boldsymbol{y}$, the inequality holds true for every constant $C$. Without loss of generality we may assume that $|\boldsymbol{y}| \geq |\boldsymbol{x}|$ and $|\boldsymbol{y}| > 0$. We can then divide the inequality (4.14) by the factor $|\boldsymbol{y}|^{\alpha}$ to get

$$\left|\frac{|\boldsymbol{x}|^{\alpha-1}\boldsymbol{x}}{|\boldsymbol{y}|^{\alpha}} - \frac{\boldsymbol{y}}{|\boldsymbol{y}|}\right| \leq C\left|\frac{\boldsymbol{x}}{|\boldsymbol{y}|} - \frac{\boldsymbol{y}}{|\boldsymbol{y}|}\right|^{\alpha}.$$

It is therefore sufficient to demonstrate the statement for $|\boldsymbol{y}| = 1$.

Let us introduce the following function

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{\left||\boldsymbol{x}|^{\alpha-1}\boldsymbol{x} - |\boldsymbol{y}|^{\alpha-1}\boldsymbol{y}\right|}{|\boldsymbol{x}-\boldsymbol{y}|^{\alpha}}.$$

The statement of the lemma will be proved once we prove that this nonnegative function has an upper bound on the set $M$

$$M = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d : |\boldsymbol{y}| = 1 \text{ and } |\boldsymbol{y}| \geq |\boldsymbol{x}|\}.$$

The function $f$ is clearly continuous for $\boldsymbol{x} \neq \boldsymbol{y}$. We will show that it is well defined and continuous for $\boldsymbol{x} = \boldsymbol{y}$ as well

$$\lim_{|\boldsymbol{\varepsilon}|\to 0} |f(\boldsymbol{y}+\boldsymbol{\varepsilon}, \boldsymbol{y}) - f(\boldsymbol{y}, \boldsymbol{y})| = \lim_{|\boldsymbol{\varepsilon}|\to 0} |f(\boldsymbol{y}+\boldsymbol{\varepsilon}, \boldsymbol{y}) - 0| = 0.$$

It holds true that

$$\begin{aligned} f(\boldsymbol{y}+\boldsymbol{\varepsilon}, \boldsymbol{y}) &= \frac{\left||\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}(\boldsymbol{y}+\boldsymbol{\varepsilon}) - |\boldsymbol{y}|^{\alpha-1}\boldsymbol{y}\right|}{|\boldsymbol{\varepsilon}|^{\alpha}} \\ &= \frac{\left|\left[|\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1} - |\boldsymbol{y}|^{\alpha-1}\right]\boldsymbol{y} + |\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}\boldsymbol{\varepsilon}\right|}{|\boldsymbol{\varepsilon}|^{\alpha}} \\ &\leq \frac{\left||\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1} - |\boldsymbol{y}|^{\alpha-1}\right||\boldsymbol{y}|}{|\boldsymbol{\varepsilon}|^{\alpha}} + |\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}|\boldsymbol{\varepsilon}|^{1-\alpha}. \end{aligned} \quad (4.15)$$

The last term on the right-hand side vanishes for $|\boldsymbol{\varepsilon}| \to \boldsymbol{0}$. The estimate of the first term falls into two cases. In the case of $|\boldsymbol{y}+\boldsymbol{\varepsilon}| \leq |\boldsymbol{y}|$, we deduce by the triangle inequality

that

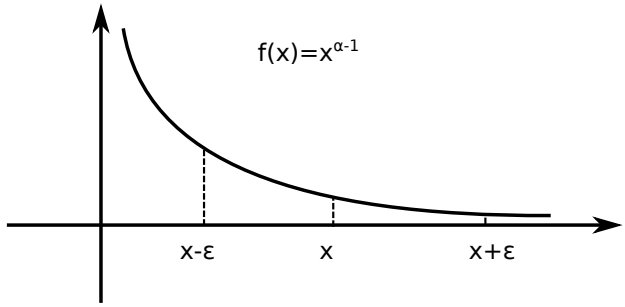
$$\frac{\left||\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}-|\boldsymbol{y}|^{\alpha-1}\right||\boldsymbol{y}|}{|\boldsymbol{\varepsilon}|^{\alpha}} = \frac{|\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}-|\boldsymbol{y}|^{\alpha-1}}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| \leq \frac{(|\boldsymbol{y}|-|\boldsymbol{\varepsilon}|)^{\alpha-1}-|\boldsymbol{y}|^{\alpha-1}}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}|.$$

The mean value theorem then leads to

$$\begin{aligned}\frac{(|\boldsymbol{y}|-|\boldsymbol{\varepsilon}|)^{\alpha-1}-|\boldsymbol{y}|^{\alpha-1}}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| &= \frac{(\alpha-1)(|\boldsymbol{y}|-\theta|\boldsymbol{\varepsilon}|)^{\alpha-2}|\boldsymbol{\varepsilon}|}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| \\ &= (\alpha-1)(|\boldsymbol{y}|-\theta|\boldsymbol{\varepsilon}|)^{\alpha-2}|\boldsymbol{\varepsilon}|^{1-\alpha}|\boldsymbol{y}| \\ &\to 0 \quad \text{for } |\boldsymbol{\varepsilon}| \to 0\end{aligned}$$

where $\theta \in (0,1)$. In the case of $|\boldsymbol{y}+\boldsymbol{\varepsilon}| > |\boldsymbol{y}|$, we analogously find that

Figure 4.3: An illustrative example




$$\begin{aligned}\frac{\left||\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}-|\boldsymbol{y}|^{\alpha-1}\right||\boldsymbol{y}|}{|\boldsymbol{\varepsilon}|^{\alpha}} &= \frac{|\boldsymbol{y}|^{\alpha-1}-|\boldsymbol{y}+\boldsymbol{\varepsilon}|^{\alpha-1}}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| \\ &\leq \frac{|\boldsymbol{y}|^{\alpha-1}-(|\boldsymbol{y}|+|\boldsymbol{\varepsilon}|)^{\alpha-1}}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| \\ &= \frac{(\alpha-1)(|\boldsymbol{y}|+\theta|\boldsymbol{\varepsilon}|)^{\alpha-2}|\varepsilon|}{|\boldsymbol{\varepsilon}|^{\alpha}}|\boldsymbol{y}| \\ &= (\alpha-1)(|\boldsymbol{y}|+\theta|\boldsymbol{\varepsilon}|)^{\alpha-2}|\boldsymbol{\varepsilon}|^{1-\alpha}|\boldsymbol{y}| \\ &\to 0 \quad \text{for } |\boldsymbol{\varepsilon}| \to 0\end{aligned}$$

and hence $f(\boldsymbol{y},\boldsymbol{y}) = 0$.

We finally use the extreme value theorem which states that a continuous function on compact set attains its maximum. The function $f$ is bounded and so there exists a constant $C$ such that the statement (4.14) is valid.

□

## 4.5 Existence and time-error estimates

We here demonstrate the existence of the solution of the problem (4.11) and derive the convergence rate of the time approximation scheme (4.12). As in the two previous chapters we use Rothe's method.

Let us recall the notation for the Rothe functions. We write $\overline{\boldsymbol{h}}_n$ for the piecewise-constant-in-time vector field

$$\begin{aligned}&\overline{\boldsymbol{h}}_n(0) = \boldsymbol{h}_0\,,\\ &\overline{\boldsymbol{h}}_n(t) = \boldsymbol{h}_i \qquad \text{for } t \in (t_{i-1}, t], \quad i = 1, \dots, n,\end{aligned}$$

and $\boldsymbol{h}_n$ for the piecewise-linear-in-time vector field

$$\begin{aligned}&\boldsymbol{h}_n(0) = \boldsymbol{h}_0\,,\\ &\boldsymbol{h}_n(t) = \boldsymbol{h}_{i-1} + (t - t_{i-1})\delta\boldsymbol{h}_i \qquad \text{for } t \in (t_{i-1}, t], \quad i = 1, \dots, n.\end{aligned}$$

The symbol $\boldsymbol{a}_n$ denotes the piecewise-linear-in-time vector field which coincides with $\boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n})$ at $t_i$, i.e.

$$\begin{aligned}&\boldsymbol{a}_n(0) = \boldsymbol{a}(\boldsymbol{h}_0 \times \boldsymbol{n})\,,\\ &\boldsymbol{a}_n(t) = \boldsymbol{a}(\boldsymbol{h}_{i-1} \times \boldsymbol{n}) + (t - t_{i-1})\delta\boldsymbol{a}(\boldsymbol{h}_i \times \boldsymbol{n}) \qquad \text{for } t \in (t_{i-1}, t], \quad i = 1, \dots, n.\end{aligned}$$

We can rewrite the approximation scheme (4.12) with this notation as follows

$$\begin{aligned}(\partial_t \boldsymbol{h}_n, \boldsymbol{\varphi}) + (\nabla \times \overline{\boldsymbol{h}}_n, \nabla \times \boldsymbol{\varphi}) + (\partial_t \boldsymbol{a}_n, \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma = 0, \quad \forall \boldsymbol{\varphi} \in \boldsymbol{V},\\ \boldsymbol{h}_n(0) = \boldsymbol{h}_0.\end{aligned} \tag{4.16}$$

The first theorem of this section claims that the solution of (4.16) converges to the limit which is the solution of the continuous problem (4.11). Let us note that the convergence in the theorem below is valid for subsequences. It is the uniqueness of the solution (see Theorem 4.1) which implies that the convergence takes place for the whole sequence.

**Theorem 4.2.** *Let $\boldsymbol{h}_0 \in \boldsymbol{V}$. Suppose that $\boldsymbol{h}_n$ and $\overline{\boldsymbol{h}}_n$ obey (4.16). Then there exists a vector field $\boldsymbol{h}$ such that*

*(i)* $\overline{\boldsymbol{h}}_n \rightharpoonup \boldsymbol{h}$ *in* $L^2((0,T), \boldsymbol{H}(\mathrm{curl}; \Omega))$,
$\boldsymbol{h}_n \rightharpoonup \boldsymbol{h}$ *in* $L^2((0,T), \boldsymbol{H}(\mathrm{curl}; \Omega))$,

*(ii)* $\boldsymbol{h}_n(t) \rightharpoonup \boldsymbol{h}(t)$ *in* $\boldsymbol{L}^2(\Omega)$ *for any* $t \in [0,T]$,
$(\partial_t \boldsymbol{h}_n, \boldsymbol{\varphi}) \to (\partial_t \boldsymbol{h}, \boldsymbol{\varphi})$ *in* $L^2((0,T), \boldsymbol{L}^2(\Omega))$,

*(iii)* $\boldsymbol{a}_n \rightharpoonup \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n})$ *in* $L_2((0,T), \boldsymbol{V}^*)$,
$\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) \rightharpoonup \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n})$ *in* $L^{\frac{1+\alpha}{\alpha}}\big((0,T), \boldsymbol{L}^{\frac{1+\alpha}{\alpha}}(\Gamma)\big)$,

*(iv)* $\boldsymbol{a}_n(t) \rightharpoonup \boldsymbol{a}(\boldsymbol{h}(t) \times \boldsymbol{n})$ *in* $\boldsymbol{V}^*$ *for any* $t \in [0,T]$,
$(\partial_t \boldsymbol{a}_n, \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma \rightharpoonup (\partial_t \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma$ *in* $L^2((0,T), \boldsymbol{V}^*)$,

*(v)* $\boldsymbol{h}$ *is a weak solution of (4.11).*

*Proof. (i)* Lemma 4.3 implies that the sequence $\overline{\boldsymbol{h}}_n$ is bounded in $L^2((0,T), \boldsymbol{H}(\mathrm{curl};\Omega))$

$$\begin{aligned}
\Big\|\overline{\boldsymbol{h}}_n\Big\|^2_{L^2((0,T),\boldsymbol{H}(\mathrm{curl};\Omega))} &= \int_0^T \big(\|\overline{\boldsymbol{h}}_n\|^2 + \|\nabla \times \overline{\boldsymbol{h}}_n\|^2\big)\,\mathrm{d}t \\
&= \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \big(\|\overline{\boldsymbol{h}}_n\|^2 + \|\nabla \times \overline{\boldsymbol{h}}_n\|^2\big)\,\mathrm{d}t \\
&= \sum_{i=1}^n \big(\|\boldsymbol{h}_i\|^2 + \|\nabla \times \boldsymbol{h}_i\|^2\big)\tau \\
&\le C.
\end{aligned}$$

This space is reflexive and hence the sequence $\overline{\boldsymbol{h}}_n$ contains a weakly convergence subsequence, i.e.

$$\overline{\boldsymbol{h}}_n \rightharpoonup \boldsymbol{h} \qquad \text{in } L^2((0,T), \boldsymbol{H}(\mathrm{curl};\Omega)).$$

That the sequence $\boldsymbol{h}_n$ has the same limit $\boldsymbol{h}$ follows from Lemma 4.3 and 4.4

$$\begin{aligned}
\int_0^T \Big\|\overline{\boldsymbol{h}}_n - \boldsymbol{h}_n\Big\|^2_{\boldsymbol{H}(\mathrm{curl};\Omega)} &= \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|(\tau - t + t_{i-1})\delta \boldsymbol{h}_n\|^2_{\boldsymbol{H}(\mathrm{curl};\Omega)}\,\mathrm{d}t \\
&= \sum_{i=1}^n 4\big(\|\boldsymbol{h}_i - \boldsymbol{h}_{i-1}\|^2 + \|\nabla \times (\boldsymbol{h}_i - \boldsymbol{h}_{i-1})\|^2\big)\tau \\
&\le \frac{C}{n} \to \infty \quad \text{for } n \to \infty.
\end{aligned}$$

*(ii)* Consider the functions $\boldsymbol{h}_n : [0,T] \to \boldsymbol{L_2}(\Omega)$, $n \in \mathbb{N}$. The sequence $\boldsymbol{h}_n$ is equibounded

$$\|\boldsymbol{h}_n(t)\| \le \|\boldsymbol{h}_{i-1}\| + \|\boldsymbol{h}_i - \boldsymbol{h}_{i-1}\| \le C \qquad \text{for any } n \in \mathbb{N}.$$

It is also uniformly equicontinuous. Indeed, we can readily establish for any $t_1, t_2 \in [0,T]$ and every $n \in \mathbb{N}$ that

$$\begin{aligned}\|\boldsymbol{h}_n(t_2) - \boldsymbol{h}_n(t_1)\| &= \left\| \int_{t_1}^{t_2} \partial_t \boldsymbol{h}_n(t)\,\mathrm{d}t \right\| \\ &\le \sqrt{\int_{t_1}^{t_2} 1^2\,\mathrm{d}t} \sqrt{\int_{t_1}^{t_2} \|\partial_t \boldsymbol{h}_n(t)\|^2\,\mathrm{d}t} \\ &\le |t_2 - t_1|^{\frac{1}{2}} \sqrt{\sum_{i=1}^{n} \|\delta \boldsymbol{h}_i\|^2 \tau} \\ &\le C|t_2 - t_1|^{\frac{1}{2}},\end{aligned}$$

where the Cauchy inequality and Lemma 4.4 have been applied. A modification of Arzela-Ascoli theorem (see [71, Lemma 1.3.10] or Lemma A.3) leads to

$$\boldsymbol{h}_n(t) \rightharpoonup \boldsymbol{h}(t) \qquad \text{in } \boldsymbol{L}^2(\Omega)$$

for any $t \in [0,T]$.

The sequence $\partial_t \boldsymbol{h}_n$ is bounded in the space $L^2((0,T), \boldsymbol{L}^2(\Omega))$ by Lemma 4.4 and so $\partial_t \boldsymbol{h}_n \rightharpoonup \partial_t \boldsymbol{h}$.

*(iii)* The sequence $\overline{\boldsymbol{h}}_n \times \boldsymbol{n}$ is bounded in the reflexive space $L^{1+\alpha}((0,T), \boldsymbol{L}^{1+\alpha}(\Gamma))$ according to Lemma 4.3. We deduce from the identity

$$\int_\Gamma |\overline{\boldsymbol{h}}_n \times \boldsymbol{n}|^{1+\alpha}\,\mathrm{d}S = \int_\Gamma |\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n})|^{(1+\alpha)/\alpha}\,\mathrm{d}S$$

that $\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) \rightharpoonup \boldsymbol{w}$ in $L^{(1+\alpha)/\alpha}((0,T), \boldsymbol{L}^{(1+\alpha)/\alpha}(\Gamma))$. The estimate

$$\begin{aligned}&\left| \int_0^T (\boldsymbol{a}_n - \boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}), \boldsymbol{\varphi} \times \boldsymbol{n})_\Gamma\,\mathrm{d}t \right| \\ \le& C\tau \int_0^T \|\partial_t \boldsymbol{a}_n\|_{\boldsymbol{V}^*} \|\boldsymbol{\varphi}\|_{\boldsymbol{V}}\,\mathrm{d}t \\ \le& \frac{C}{n} \sqrt{\int_0^T \|\boldsymbol{\varphi}\|_{\boldsymbol{V}}^2\,\mathrm{d}t}\end{aligned} \tag{4.17}$$

implies that the sequences $\boldsymbol{a}_n$ and $\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n})$ have the same weak limit $\boldsymbol{w}$ in the space $L_2((0,T), \boldsymbol{V}^*)$.

We use now the Browder-Minty trick (see [45] or Lemma A.1) to demonstrate that $\boldsymbol{n} \times \boldsymbol{w} = \boldsymbol{n} \times \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n})$. Here follow two relations which will be used later. We first

integrate (4.16) in time and set $\boldsymbol{\varphi} = \overline{\boldsymbol{h}}_n$. Integrating in time once again, we obtain

$$\begin{aligned}\int_0^T (\boldsymbol{h}_n, \overline{\boldsymbol{h}}_n)\,\mathrm{d}t + \int_0^T \left( \int_0^t \nabla \times \overline{\boldsymbol{h}}_n \,\mathrm{d}s, \nabla \times \overline{\boldsymbol{h}}_n \right) \mathrm{d}t + \int_0^T (\boldsymbol{a}_n, \overline{\boldsymbol{h}}_n \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \\ = \int_0^T (\boldsymbol{h}_0, \boldsymbol{h}_n)\,\mathrm{d}t + \int_0^T (\boldsymbol{a}(\boldsymbol{h}_0 \times \boldsymbol{n}), \overline{\boldsymbol{h}}_n \times \boldsymbol{n})_\Gamma \,\mathrm{d}t\,.\end{aligned} \tag{4.18}$$

We then integrate (4.16) in time and set $\boldsymbol{\varphi} = \boldsymbol{h}$. Integration in time once again and taking $n \to \infty$ yield

$$\begin{aligned}\int_0^T (\boldsymbol{h}, \boldsymbol{h})\,\mathrm{d}t + \int_0^T \left( \int_0^t \nabla \times \boldsymbol{h} \,\mathrm{d}s, \nabla \times \boldsymbol{h} \right) \mathrm{d}t + \int_0^T (\boldsymbol{w}, \boldsymbol{h} \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \\ = \int_0^T (\boldsymbol{h}_0, \boldsymbol{h})\,\mathrm{d}t + \int_0^T (\boldsymbol{a}(\boldsymbol{h}_0 \times \boldsymbol{n}), \boldsymbol{h} \times \boldsymbol{n})_\Gamma \,\mathrm{d}t\,.\end{aligned} \tag{4.19}$$

It holds by the monotonicity argument that

$$\int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{u} \times \boldsymbol{n}), \overline{\boldsymbol{h}}_n \times \boldsymbol{n} - \boldsymbol{u} \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \geq 0$$

for any $\boldsymbol{u}$ in $L^{1+\alpha}((0,T), \boldsymbol{L}^{1+\alpha}(\Gamma))$. Taking the limit for $n \to \infty$ leads to

$$\int_0^T (-\boldsymbol{a}(\boldsymbol{u} \times \boldsymbol{n}), \overline{\boldsymbol{h}}_n \times \boldsymbol{n} - \boldsymbol{u} \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \to \int_0^T (-\boldsymbol{a}(\boldsymbol{u} \times \boldsymbol{n}), \boldsymbol{h} \times \boldsymbol{n} - \boldsymbol{u} \times \boldsymbol{n})_\Gamma \,\mathrm{d}t\,.$$

We consecutively deduce on account of the previous results that

$$
\begin{aligned}
&\lim_{n\to\infty}\int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n\times\boldsymbol{n}),\overline{\boldsymbol{h}}_n\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\\
&\overset{(4.17)}{=}\lim_{n\to\infty}\int_0^T (\boldsymbol{a}_n,\overline{\boldsymbol{h}}_n\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\\
&\overset{(4.18)}{=}\lim_{n\to\infty}\left[\int_0^T(\boldsymbol{h}_0,\overline{\boldsymbol{h}}_n)\,\mathrm{d}t+\int_0^T(\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\overline{\boldsymbol{h}}_n\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\right.\\
&\quad\left.-\int_0^T(\boldsymbol{h}_n,\overline{\boldsymbol{h}}_n)\,\mathrm{d}t-\int_0^T\left(\int_0^t\nabla\times\overline{\boldsymbol{h}}_n\,\mathrm{d}s,\nabla\times\overline{\boldsymbol{h}}_n\right)\mathrm{d}t\right]\\
&=\int_0^T(\boldsymbol{h}_0,\boldsymbol{h})\,\mathrm{d}t+\int_0^T(\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\boldsymbol{h}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\\
&\quad-\lim_{n\to\infty}\left[\int_0^T(\boldsymbol{h}_n,\overline{\boldsymbol{h}}_n)\,\mathrm{d}t+\frac{1}{2}\left\|\int_0^T\nabla\times\overline{\boldsymbol{h}}_n\,\mathrm{d}t\right\|^2\right]\\
&\overset{(*)}{\le}\int_0^T(\boldsymbol{h}_0,\boldsymbol{h})\,\mathrm{d}t+\int_0^T(\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\boldsymbol{h}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\\
&\quad-\int_0^T(\boldsymbol{h},\boldsymbol{h})\,\mathrm{d}t-\frac{1}{2}\left\|\int_0^T\nabla\times\boldsymbol{h}\,\mathrm{d}t\right\|^2\\
&\overset{(4.19)}{=}\int_0^T(\boldsymbol{w},\boldsymbol{h}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\,.
\end{aligned}
$$

The inequality $(*)$ is justified by the relation $\boldsymbol{h}_n(t)=\overline{\boldsymbol{h}}_n(t)-\partial_t\boldsymbol{h}_n(t-t_{i-1})$ and by the weak lower semicontinuity of the norm. Collecting the results, we obtain

$$
\int_0^T(\boldsymbol{w}-\boldsymbol{a}(\boldsymbol{u}\times\boldsymbol{n}),\boldsymbol{h}\times\boldsymbol{n}-\boldsymbol{u}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\ge 0\,.
$$

The substitution $\boldsymbol{u}=\boldsymbol{h}+\varepsilon\boldsymbol{v}$ for any $\boldsymbol{v}\in L^{1+\alpha}((0,T),\boldsymbol{L}^{1+\alpha}(\Gamma))$ and $\varepsilon>0$ gives

$$
\int_0^T(\boldsymbol{w}-\boldsymbol{a}((\boldsymbol{h}+\varepsilon\boldsymbol{v})\times\boldsymbol{n}),\boldsymbol{v}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\le 0
$$

and so it holds for the limit case $\varepsilon\to 0$ that

$$
\int_0^T(\boldsymbol{w}-\boldsymbol{a}((\boldsymbol{h}\times\boldsymbol{n})),\boldsymbol{v}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\le 0.
$$

The above inequality is valid for both $\boldsymbol{v}$ and $-\boldsymbol{v}$ which forces

$$
\int_0^T(\boldsymbol{w}-\boldsymbol{a}((\boldsymbol{h}\times\boldsymbol{n})),\boldsymbol{v}\times\boldsymbol{n})_\Gamma\,\mathrm{d}t=0
$$

for any $\boldsymbol{v} \in L^{1+\alpha}((0,T),\boldsymbol{L}^{1+\alpha}(\Gamma))$. Therefore

$$\boldsymbol{n}\times\boldsymbol{w} = \boldsymbol{n}\times\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}) \quad \text{a.e. in} \quad \Gamma\times(0,T)\,.$$

*(iv)* Lemma (4.5) implies that

$$\|\boldsymbol{a}_n(t)\|_{\boldsymbol{V}^*} = \left\|\boldsymbol{a}_n(0) + \int_0^t \partial_s \boldsymbol{a}_n \,\mathrm{d}s\right\|_{\boldsymbol{V}^*} \leq C$$

for any $t\in[0,T]$. It moreover holds that

$$\|\boldsymbol{a}_n(t_2)-\boldsymbol{a}_n(t_1)\|_{\boldsymbol{V}^*} = \left\|\int_{t_1}^{t_2} \partial_t \boldsymbol{a}_n \,\mathrm{d}t\right\|_{\boldsymbol{V}^*} \leq C|t_2-t_1|,$$

and so we have according [71, Lemma 1.3.10] and $(iii)$ that

$$\boldsymbol{a}_n(t) \rightharpoonup \boldsymbol{a}(\boldsymbol{h}(t)\times\boldsymbol{n}) \quad \text{ in } \boldsymbol{V}^* \text{ for any } t\in[0,T].$$

Lemma (4.5) yields $\partial_t \boldsymbol{a}_n \rightharpoonup \boldsymbol{z}$ by standard argument. Letting $n\to\infty$ in the identity

$$(\boldsymbol{a}_n(t)-\boldsymbol{a}_n(0),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma = \int_0^t (\partial_s\boldsymbol{a}_n,\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma \,\mathrm{d}s,$$

we obtain

$$(\boldsymbol{a}(\boldsymbol{h}(t)\times\boldsymbol{n}),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma = (\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma + \int_0^t (\boldsymbol{z},\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma \,\mathrm{d}s.$$

Since it is valid for any $t\in[0,T]$, we have

$$\boldsymbol{n}\times\boldsymbol{z} = \boldsymbol{n}\times\partial_t\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}).$$

*(v)* Integration of (4.16) in time gives

$$\begin{aligned}(\boldsymbol{h}_n(t),\boldsymbol{\varphi}) - (\boldsymbol{h}_0,\boldsymbol{\varphi}) + \int_0^t (\nabla\times\overline{\boldsymbol{h}}_n,\nabla\times\boldsymbol{\varphi})\,\mathrm{d}s + (\boldsymbol{a}_n(t),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma \\ - (\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma = 0.\end{aligned}$$

We make use of *(i)*-*(iv)* to take the limit $n\to\infty$. The above formula becomes

$$\begin{aligned}(\boldsymbol{h}(t),\boldsymbol{\varphi}) - (\boldsymbol{h}_0,\boldsymbol{\varphi}) + \int_0^t (\nabla\times\boldsymbol{h},\nabla\times\boldsymbol{\varphi})\,\mathrm{d}s + (\boldsymbol{a}(\boldsymbol{h}(t)\times\boldsymbol{n}),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma \\ - (\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),\boldsymbol{\varphi}\times\boldsymbol{n})_\Gamma = 0,\end{aligned}$$

and differentiation with the respect to time concludes the proof. □

The second (and last) theorem of this section states error estimates for the time discretization scheme.

**Theorem 4.3.** *Let $\boldsymbol{h}_0 \in \boldsymbol{V}$. Suppose that $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}_n$ are the solutions of (4.11) and (4.16) respectively. Then there exists a positive constant $C$ such that*

$$\int_0^T \|\overline{\boldsymbol{h}}_n - \boldsymbol{h}\|^2 \,\mathrm{d}t + \left\| \int_0^T \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \right\|^2 \mathrm{d}t + \int_0^T \int_\Gamma (|\overline{\boldsymbol{h}}_n \times \boldsymbol{n}|^{(\alpha+1)/2} - |\boldsymbol{h} \times \boldsymbol{n}|^{(\alpha+1)/2})^2 \,\mathrm{d}S \,\mathrm{d}t \le C\tau.$$

*Proof.* We first subtract (4.11) from (4.16) and integrate in time. Putting $\boldsymbol{\varphi} = \overline{\boldsymbol{h}}_n - \boldsymbol{h}$ and integrating again in time, we get

$$\begin{aligned} &\int_0^T (\overline{\boldsymbol{h}}_n - \boldsymbol{h}, \overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}t \\ &+ \int_0^T \left( \int_0^t \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}s, \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \right) \mathrm{d}t \\ &+ \int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}), (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \\ &= \int_0^T (\overline{\boldsymbol{h}}_n - \boldsymbol{h}_n, \overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}t + \int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) - \boldsymbol{a}_n, (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \times \boldsymbol{n})_\Gamma \,\mathrm{d}t. \end{aligned} \tag{4.20}$$

The second term on the LHS of (4.20) can be rewritten in the form

$$\int_0^T \left( \int_0^t \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}s, \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h})(t) \right) \mathrm{d}t = \frac{1}{2} \left\| \int_0^T \nabla \times (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}t \right\|^2.$$

In view of the following algebraic inequality

$$4ab\big(y^{(a+b)/2} - z^{(a+b)/2}\big) \le (a+b)^2 (y^a - z^a)(y^b - z^b)$$

which is valid for any $a, b, y, z \ge 0$, we can derive by the Cauchy inequality that

$$\begin{aligned} (\boldsymbol{y} - \boldsymbol{z}) \cdot (|\boldsymbol{y}|^{\alpha-1}\boldsymbol{y} - |\boldsymbol{z}|^{\alpha-1}\boldsymbol{z}) &= |\boldsymbol{y}|^{\alpha+1} + |\boldsymbol{z}|^{\alpha+1} - |\boldsymbol{z}|^{\alpha-1}\boldsymbol{z} \cdot \boldsymbol{y} - |\boldsymbol{y}|^{\alpha-1}\boldsymbol{z} \cdot \boldsymbol{y} \\ &\ge |\boldsymbol{y}|^{\alpha+1} + |\boldsymbol{z}|^{\alpha+1} - |\boldsymbol{z}|^{\alpha}|\boldsymbol{y}| - |\boldsymbol{y}|^{\alpha}|\boldsymbol{z}| \\ &= (|\boldsymbol{y}|^{\alpha} - |\boldsymbol{z}|^{\alpha})(|\boldsymbol{y}| - |\boldsymbol{z}|) \\ &\ge \frac{4\alpha}{(\alpha+1)^2} (|\boldsymbol{y}|^{(\alpha+1)/2)} - |\boldsymbol{z}|^{(\alpha+1)/2})^2. \end{aligned}$$

This implies for the third term on LHS of (4.20) that

$$\begin{aligned}&\int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h} \times \boldsymbol{n}), (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \\ \geq &\frac{4\alpha}{(\alpha+1)^2} \int_0^T \int_\Gamma [|\overline{\boldsymbol{h}}_n \times \boldsymbol{n}|^{(\alpha+1)/2} - |\boldsymbol{h} \times \boldsymbol{n}|^{(\alpha+1)/2}]^2 \,\mathrm{d}S \,\mathrm{d}t\,.\end{aligned}$$

It remains to estimate the RHS of (4.20). It follows from the a priori estimates from Lemma 4.4 and Lemma 4.5 that

$$\begin{aligned}&\left| \int_0^T (\overline{\boldsymbol{h}}_n(t) - \boldsymbol{h}_n(t), \overline{\boldsymbol{h}}_n - \boldsymbol{h}) \,\mathrm{d}t + \int_0^T (\boldsymbol{a}(\overline{\boldsymbol{h}}_n \times \boldsymbol{n}) - \boldsymbol{a}_n, (\overline{\boldsymbol{h}}_n - \boldsymbol{h}) \times \boldsymbol{n})_\Gamma \,\mathrm{d}t \right| \\ \leq &C\tau \left( \int_0^T \|\partial_t \boldsymbol{h}_n\| \|\overline{\boldsymbol{h}}_n - \boldsymbol{h}\| \,\mathrm{d}t + \int_0^T \|\partial_t \boldsymbol{a}_n\|_{\boldsymbol{V}^*} \|\overline{\boldsymbol{h}}_n - \boldsymbol{h}\|_{\boldsymbol{V}} \,\mathrm{d}t \right) \\ \leq &C\tau.\end{aligned}$$

□

## 4.6 Full discretization

The aim of this section is to investigate a full discretization of the problem (4.11). We reformulate in particular the time discretization results (4.16) to the space discretized version.

Let $\boldsymbol{V}^h$ be a finite dimensional subspace of $\boldsymbol{V}$ and the mapping $\boldsymbol{r}_h : \boldsymbol{V} \to \boldsymbol{V}^h$ be a linear bounded projection operator onto it. The full discretized version of the original problem (4.11) is to find $\boldsymbol{u}_i^h \in \boldsymbol{V}^h$ such that the equation

$$\begin{aligned}(\delta \boldsymbol{u}_i^h, \boldsymbol{\varphi}^h) + (\nabla \times \boldsymbol{u}_i^h, \nabla \times \boldsymbol{\varphi}^h) + (\delta \boldsymbol{a}(\boldsymbol{u}_i^h \times \boldsymbol{n}), \boldsymbol{\varphi}^h \times \boldsymbol{n})_\Gamma = 0, \qquad (4.21)\\ \boldsymbol{u}_0^h = \boldsymbol{r}_h \boldsymbol{h}_0\end{aligned}$$

holds for any $i = 1, \ldots, n$ and $\boldsymbol{\varphi}^h \in \boldsymbol{V}^h$. This problem admits a unique solution $\boldsymbol{u}_i^h$, similarly to (4.12).

**Lemma 4.7.** *Suppose $\boldsymbol{h}_0 \in \boldsymbol{V}$, then there exists the uniquely determined $\boldsymbol{u}_i^h \in \boldsymbol{V}^h$ solving (4.21) for any index $i = 1, \ldots, n$.*

*Proof.* The proof follows the same line as in Lemma (4.1). Instead of the mapping $A$, we consider its finite dimensional approximation $A^h : \boldsymbol{V}^h \to (\boldsymbol{V}^h)^*$

$$\langle A^h(\boldsymbol{u}^h), \boldsymbol{\varphi}^h \rangle = \left( \frac{\boldsymbol{u}^h}{\tau}, \boldsymbol{\varphi}^h \right) + (\nabla \times \boldsymbol{u}^h, \nabla \times \boldsymbol{\varphi}^h) + \left( \frac{\boldsymbol{a}(\boldsymbol{u}^h \times \boldsymbol{n})}{\tau}, \boldsymbol{\varphi}^h \times \nu \right)_\Gamma$$

where $\boldsymbol{\varphi}^h \in \boldsymbol{V}^h$. □

The next two lemmas state a priori estimates. We omit the proofs, since they are similar to the ones of Lemma 4.3 and 4.4.

**Lemma 4.8.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{V}$. Then there exists a positive constant $C$ such that*

$$\|\boldsymbol{u}_j^h\|^2 + \sum_{i=1}^{j} \|\boldsymbol{u}_i^h - \boldsymbol{u}_{i-1}^h\|^2 + \sum_{i=1}^{j} \|\nabla \times \boldsymbol{u}_i^h\|^2 \tau + \|\boldsymbol{u}_j^h \times \boldsymbol{n}\|_{\boldsymbol{L}_{1+\alpha}^{1+\alpha}(\Gamma)} \leq C$$

*for any $j = 1, \ldots, n$.*

**Lemma 4.9.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{V}$. Then there exists a positive constant $C$ such that*

$$\sum_{i=1}^{j} \|\delta \boldsymbol{u}_i^h\|^2 \tau + \|\nabla \times \boldsymbol{u}_j^h\|^2 + \sum_{i=1}^{j} \|\nabla \times (\boldsymbol{u}_i^h - \boldsymbol{u}_{i-1}^h)\|^2 \leq C$$

*for any $j = 1, \ldots, n$.*

The full discretized system (4.21) can be rewritten by Rothe's notation as follows

$$\begin{aligned} (\partial_t \boldsymbol{u}_n^h, \boldsymbol{\varphi}^h) + (\nabla \times \overline{\boldsymbol{u}}_n^h, \nabla \times \boldsymbol{\varphi})^h + (\partial_t \boldsymbol{a}_n^h, \boldsymbol{\varphi}^h \times \boldsymbol{n})_\Gamma = 0, \\ \boldsymbol{u}_n^h(0) = \boldsymbol{r}_h \boldsymbol{h}_0. \end{aligned} \tag{4.22}$$

The next theorem establishes an error estimate for the full discretization.

**Theorem 4.4.** *Let $\boldsymbol{h}_0 \in \boldsymbol{V}$. Suppose that $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}_n$ are the solutions of (4.11) and (4.21) respectively. Then the error estimate*

$$\begin{aligned} &\int_0^T \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \,\mathrm{d}t + \left\| \int_0^T \nabla \times (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h}) \,\mathrm{d}t \right\|^2 \\ &+ \int_0^T \int_\Gamma \left( |\overline{\boldsymbol{u}}_n^h \times \boldsymbol{n}|^{(\alpha+1)/2} - |\boldsymbol{h} \times \boldsymbol{n}|^{(\alpha+1)/2} \right)^2 \mathrm{d}S \,\mathrm{d}t \\ &\leq \\ &C \Bigg( \tau + \|\boldsymbol{r}_h \boldsymbol{h}_0 - \boldsymbol{h}_0\|^2 + \|(\boldsymbol{h}_0 - \boldsymbol{r}_h \boldsymbol{h}_0) \times \boldsymbol{n}\|_{\boldsymbol{L}_{1+\alpha}(\Gamma)}^{\alpha} \\ &\qquad + \|(\boldsymbol{h} - \boldsymbol{r}_h \boldsymbol{h}) \times \boldsymbol{n}\|_{L^{1+\alpha}\left((0,T), \boldsymbol{L}^{1+\alpha}(\Gamma)\right)} \\ &+ \int_0^T \|\nabla \times (\boldsymbol{h} - \boldsymbol{r}_h \boldsymbol{h})\|^2 \,\mathrm{d}t + \int_0^T \|\boldsymbol{h} - \boldsymbol{r}_h \boldsymbol{h}\|^2 \,\mathrm{d}t \Bigg) \end{aligned}$$

*holds.*

*Proof.* We subtract (4.11) from (4.22) and integrate in time. Taking $\boldsymbol{\varphi} = \boldsymbol{\varphi}^h = \overline{\boldsymbol{u}}_n^h(t) - \boldsymbol{r}_h\boldsymbol{h}(t)$ and integration in time again over $(0,\eta)$ for $\eta \in [0,T]$ give

$$\begin{aligned}
&\int_0^\eta (\boldsymbol{u}_n^h - \boldsymbol{h}, \overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\,\mathrm{d}t \\
&+ \int_0^\eta \left( \int_0^t \nabla \times (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}s, \nabla \times (\overline{\boldsymbol{u}}_n^h(t) - \boldsymbol{r}_h\boldsymbol{h}(t)) \right) \mathrm{d}t \\
&\quad + \int_0^\eta \left(\boldsymbol{a}_n^h - \boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t \\
&= \\
&\int_0^\eta (\boldsymbol{r}_h\boldsymbol{h}_0 - \boldsymbol{h}_0, \overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\,\mathrm{d}t \\
&+ \int_0^\eta \left(\boldsymbol{a}(\boldsymbol{r}_h\boldsymbol{h}_0\times\boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t\,.
\end{aligned}$$

We rearrange the terms to obtain

$$\begin{aligned}
&\int_0^\eta \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \mathrm{d}t + \int_0^\eta \left( \int_0^t \nabla \times (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}s, \nabla \times (\overline{\boldsymbol{u}}_n^h(t) - \boldsymbol{h}(t)) \right) \mathrm{d}t \\
&\quad + \int_0^\eta \left(\boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t \\
&= \\
&\quad - \int_0^\eta (\boldsymbol{u}_n^h - \boldsymbol{h}, (\overline{\boldsymbol{u}}_n^h - \boldsymbol{u}_n^h) + (\boldsymbol{h} - \boldsymbol{r}_h\boldsymbol{h}))\,\mathrm{d}t \qquad (4.23) \\
&- \int_0^\eta \left( \int_0^t \nabla \times (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}s, \nabla \times (\boldsymbol{h}(t) - \boldsymbol{r}_h\boldsymbol{h}(t)) \right) \mathrm{d}t \\
&\quad - \int_0^\eta \left(\boldsymbol{a}_n^h - \boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t \\
&\quad - \int_0^\eta \left(\boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}), (\boldsymbol{h} - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t \\
&\quad + \int_0^\eta (\boldsymbol{r}_h\boldsymbol{h}_0 - \boldsymbol{h}_0, \overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\,\mathrm{d}t \\
&+ \int_0^\eta \left(\boldsymbol{a}(\boldsymbol{r}_h\boldsymbol{h}_0\times\boldsymbol{n}) - \boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t\,.
\end{aligned}$$

Note that all the terms on the LHS are non-negative. We can analogously to Theorem

4.3 derive the following lower bound for the left-hand side

$$\int_0^\eta \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \,\mathrm{d}t + \frac{1}{2}\left\|\int_0^\eta \nabla\times(\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}t\right\|^2 + \frac{4\alpha}{(1+\alpha)^2}\int_0^\eta\int_\Gamma (|\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}|^{(\alpha+1)/2} - |\boldsymbol{h}\times\boldsymbol{n}|^{(\alpha+1)/2})^2\,\mathrm{d}S\,\mathrm{d}t.$$

We examine the RHS of (4.23) term by term. If follows from Lemma (4.3) that

$$\begin{aligned}\left|\int_0^\eta (\boldsymbol{u}_n^h - \boldsymbol{h}, \overline{\boldsymbol{u}}_n^h - \boldsymbol{u}_n^h)\,\mathrm{d}t\right| &\le \varepsilon\int_0^\eta \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \,\mathrm{d}t + C_\varepsilon\int_0^\eta \left\|\overline{\boldsymbol{u}}_n^h - \boldsymbol{u}_n^h\right\|^2 \,\mathrm{d}t\\ &\le \varepsilon\int_0^\eta \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \,\mathrm{d}t + C_\varepsilon\tau^2.\end{aligned}$$

It holds that

$$\left|\int_0^\eta (\boldsymbol{u}_n^h - \boldsymbol{h}, \boldsymbol{h} - \boldsymbol{r}_h\boldsymbol{h})\,\mathrm{d}t\right| \le \varepsilon\int_0^\eta \left\|\boldsymbol{u}_n^h - \boldsymbol{h}\right\|^2 \,\mathrm{d}t + C_\varepsilon\int_0^\eta \left\|\boldsymbol{h} - \boldsymbol{r}_h\boldsymbol{h}\right\|^2 \,\mathrm{d}t$$

and

$$\begin{aligned}&\left|\int_0^\eta \left(\int_0^t \nabla\times(\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}s, \nabla\times(\boldsymbol{h}(t) - \boldsymbol{r}_h\boldsymbol{h}(t))\right)\mathrm{d}t\right|\\ &\le \int_0^\eta \left\|\int_0^t \nabla\times(\overline{\boldsymbol{u}}_n^h - \boldsymbol{h})\,\mathrm{d}s\right\|^2 \,\mathrm{d}t + \int_0^\eta \left\|\nabla\times(\boldsymbol{h} - \boldsymbol{r}_h\boldsymbol{h})\right\|^2 \,\mathrm{d}t.\end{aligned}$$

We use (4.16) to obtain

$$\begin{aligned}(\boldsymbol{a}_n^h - \boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}), \boldsymbol{\varphi}^h\times\boldsymbol{n})_\Gamma &= \left(\int_{t_i}^t \partial_s\boldsymbol{a}_n^h\,\mathrm{d}s, \boldsymbol{\varphi}^h\right)_\Gamma\\ &= -\left(\int_{t_i}^t \partial_s\boldsymbol{u}_n^h\,\mathrm{d}s, \boldsymbol{\varphi}^h\right) - \left(\int_{t_i}^t \nabla\times\overline{\boldsymbol{u}}_n^h\,\mathrm{d}s, \nabla\times\boldsymbol{\varphi}^h\right)\\ &\le \tau\left\|\partial_t\boldsymbol{u}_n^h\right\|\left\|\boldsymbol{\varphi}^h\right\| + \tau\left\|\nabla\times\overline{\boldsymbol{u}}_n^h\right\|\left\|\nabla\times\boldsymbol{\varphi}^h\right\|\end{aligned}$$

for $t\in(t_{i-1},t_i]$. We deduce then by Lemma 4.8 and Theorem 4.2 that

$$\left|\int_0^\eta (\boldsymbol{a}_n^h - \boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}), (\overline{\boldsymbol{u}}_n^h - \boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n})_\Gamma\,\mathrm{d}t\right| \le C\tau.$$

Applying the Hölder inequality yields

$$\begin{aligned}
&\left|\int_0^\eta \left(\boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n})-\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}),(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma \mathrm{d}t\right| \\
\leq&\int_0^\eta \left\|\boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n})-\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n})\right\|_{\boldsymbol{L}_{\frac{1+\alpha}{\alpha}}(\Gamma)}\left\|(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right\|_{\boldsymbol{L}_{1+\alpha}(\Gamma)}\,\mathrm{d}t \\
\leq&\left(\int_0^\eta \left\|\boldsymbol{a}(\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n})-\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n})\right\|^{\frac{1+\alpha}{\alpha}}_{\boldsymbol{L}_{\frac{1+\alpha}{\alpha}}(\Gamma)}\,\mathrm{d}t\right)^{\frac{\alpha}{1+\alpha}}\left(\int_0^\eta \left\|(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right\|^{1+\alpha}_{\boldsymbol{L}_{1+\alpha}(\Gamma)}\,\mathrm{d}t\right)^{\frac{1}{1+\alpha}} \\
\leq&C\left\|(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right\|_{L_{1+\alpha}((0,T),\boldsymbol{L}_{1+\alpha}(\Gamma))}.
\end{aligned}$$

For the terms coming from the initial value condition, it is easy to see that

$$\begin{aligned}
\int_0^\eta(\boldsymbol{r}_h\boldsymbol{h}_0-\boldsymbol{h}_0,\overline{\boldsymbol{u}}_n^h-\boldsymbol{r}_h\boldsymbol{h})\,\mathrm{d}t&\leq C_\varepsilon\left\|\boldsymbol{r}_h\boldsymbol{h}_0-\boldsymbol{h}_0\right\|^2+\varepsilon\int_0^\eta\left\|\overline{\boldsymbol{u}}_n^h-\boldsymbol{r}_h\boldsymbol{h}\right\|^2\,\mathrm{d}t \\
&\leq C_\varepsilon\left\|\boldsymbol{r}_h\boldsymbol{h}_0-\boldsymbol{h}_0\right\|^2+\varepsilon\int_0^\eta\left\|\boldsymbol{u}_n^h-\boldsymbol{r}_h\boldsymbol{h}\right\|^2\,\mathrm{d}t \\
&\quad+C_\varepsilon\tau^2.
\end{aligned}$$

Recalling Lemma 4.6 and the Hölder inequality, we deduce that

$$\begin{aligned}
&\left|\int_0^\eta\left(\boldsymbol{a}(\boldsymbol{r}_h\boldsymbol{h}_0\times\boldsymbol{n})-\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n}),(\overline{\boldsymbol{u}}_n^h-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right)_\Gamma\mathrm{d}t\right| \\
\leq&\int_0^\eta\int_\Gamma|\boldsymbol{a}(\boldsymbol{r}_h\boldsymbol{h}_0\times\boldsymbol{n})-\boldsymbol{a}(\boldsymbol{h}_0\times\boldsymbol{n})||(\overline{\boldsymbol{u}}_n^h-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}|\,\mathrm{d}S\,\mathrm{d}t \\
\leq&C\int_0^\eta\int_\Gamma|(\boldsymbol{h}_0-\boldsymbol{r}_h\boldsymbol{h}_0)\times\boldsymbol{n}|^\alpha|(\overline{\boldsymbol{u}}_n^h-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}|\,\mathrm{d}S\,\mathrm{d}t \\
\leq&C\left\|(\boldsymbol{h}_0-\boldsymbol{r}_h\boldsymbol{h}_0)\times\boldsymbol{n}\right\|^\alpha_{\boldsymbol{L}_{1+\alpha}(\Gamma)}.
\end{aligned}$$

We collect all the results to find that for $\tau<1$

$$\begin{aligned}
&(1-\varepsilon)\int_0^\eta\left\|\boldsymbol{u}_n^h-\boldsymbol{h}\right\|^2\,\mathrm{d}t+\frac{1}{2}\left\|\int_0^\eta(\nabla\times(\overline{\boldsymbol{u}}_n^h-\boldsymbol{h})\,\mathrm{d}t\right\|^2 \\
&+\frac{4\alpha}{(1+\alpha)^2}\int_0^\eta\int_\Gamma(|\overline{\boldsymbol{u}}_n^h\times\boldsymbol{n}|^{(\alpha+1)/2}-|\boldsymbol{h}\times\boldsymbol{n}|^{(\alpha+1)/2})^2\,\mathrm{d}S\,\mathrm{d}t \\
\leq\;&C_\varepsilon\Bigg(\tau+\left\|\boldsymbol{r}_h\boldsymbol{h}_0-\boldsymbol{h}_0\right\|^2+\left\|(\boldsymbol{h}_0-\boldsymbol{r}_h\boldsymbol{h}_0)\times\boldsymbol{n}\right\|^\alpha_{\boldsymbol{L}_{1+\alpha}(\Gamma)} \\
&+\left\|(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\times\boldsymbol{n}\right\|_{L_{1+\alpha}((0,T),\boldsymbol{L}_{1+\alpha}(\Gamma))}+\int_0^\eta\left\|\nabla\times(\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h})\right\|^2\,\mathrm{d}t \\
&+\int_0^\eta\left\|\boldsymbol{h}-\boldsymbol{r}_h\boldsymbol{h}\right\|^2\,\mathrm{d}t+\int_0^\eta\left\|\int_0^t\nabla\times(\overline{\boldsymbol{u}}_n^h-\boldsymbol{h})\,\mathrm{d}s\right\|^2\,\mathrm{d}t\Bigg).
\end{aligned}$$

This estimate is valid for any $\eta \in (0,T)$. We fix a sufficiently small positive $\varepsilon$ and apply the Gronwall lemma to conclude the proof. □

We are furthermore interested in the convergence estimates for a particular choice of the space discretization. Unlike in the previous chapter, we will state an assertion for the first-order edge elements (see Example A.2), which are curl conforming, [84, Theorem 5.37]. Let $T_h$ be a regular family of triangulations with the mesh parameter $h$ (see Definition A.21. Denote by $\boldsymbol{r}_h$ the global interpolant on the space spanned by the first order edge finite elements. It is analogous to $\pi_h$ in (A.6). The interpolant $\boldsymbol{r}_h$ is not well defined for a general function in $\boldsymbol{H}(\mathrm{curl};\Omega)$, but at least for any function from its subspace $\boldsymbol{H}^s(\Omega)$ for $s > 1/2$ according to [Theorem 5.38, ibid]. If $\boldsymbol{u} \in \boldsymbol{H}^s(\Omega)$ and $\nabla \times \boldsymbol{u} \in \boldsymbol{H}^s(\Omega)$ for the exponent $s \in (1/2, 1]$, then it follows from [Theorem 5.41, ibid] that

$$\|\boldsymbol{r}_h\boldsymbol{u} - \boldsymbol{u}\| + \|\nabla \times (\boldsymbol{r}_h\boldsymbol{u} - \boldsymbol{u})\| \le Ch^s \left( \|\boldsymbol{u}\|_{\boldsymbol{H}^s(\Omega)} + \|\nabla \times \boldsymbol{u}\|_{\boldsymbol{H}^s(\Omega)} \right).$$

The boundary trace belongs to $\boldsymbol{L}^2(\Gamma)$ according to the trace theorem (see [12, Theorem 5.20]). It satisfies the estimate (compare with [84, Theorem 5.52])

$$\|(\boldsymbol{r}_h\boldsymbol{u} - \boldsymbol{u}) \times \boldsymbol{n}\|_{\boldsymbol{L}^2(\Gamma)} \le Ch^{s-1/2} \left( \|\boldsymbol{u}\|_{\boldsymbol{H}^s(\Omega)} + \|\nabla \times \boldsymbol{u}\|_{\boldsymbol{H}^s(\Omega)} \right)$$

and by Sobolev imbeddings

$$\|(\boldsymbol{r}_h\boldsymbol{u} - \boldsymbol{u}) \times \boldsymbol{n}\|_{\boldsymbol{L}^{1+\alpha}(\Gamma)} \le C \|(\boldsymbol{r}_h\boldsymbol{u} - \boldsymbol{u}) \times \boldsymbol{n}\|_{\boldsymbol{L}^2(\Gamma)}.$$

There follows a corollary of Theorem 4.4 for first-order edge elements.

**Theorem 4.5.** *Assume $\boldsymbol{h}_0 \in \boldsymbol{H}^s(\Omega)$. Let $\boldsymbol{h}$ and $\nabla \times \boldsymbol{h}$ be from $L_2((0,T), \boldsymbol{H}^s(\Omega))$ for some $s \in (\frac{1}{2}, 1]$. Then*

$$\begin{aligned} &\int_0^\eta \left\| \boldsymbol{u}_n^h - \boldsymbol{h} \right\|^2 \mathrm{d}t + \left\| \int_0^\eta \nabla \times (\overline{\boldsymbol{u}}_n^h - \boldsymbol{h}) \, \mathrm{d}t \right\|^2 \\ &+ \int_0^\eta \int_\Gamma (|\overline{\boldsymbol{u}}_n^h \times \boldsymbol{n}|^{(\alpha+1)/2} - |\boldsymbol{h} \times \boldsymbol{n}|^{(\alpha+1)/2})^2 \, \mathrm{d}S \, \mathrm{d}t \\ &\le C \left( \tau + h^{2s} + h^{s-1/2} + h^{\alpha(s-1/2)} \right). \end{aligned}$$

*Proof.* We obtain directly from the above considerations that

$$\|\boldsymbol{r}_h\boldsymbol{h}_0 - \boldsymbol{h}_0\|^2 + \int_0^\eta \|\nabla \times (\boldsymbol{r}_h\boldsymbol{h} - \boldsymbol{h})\|^2 \, \mathrm{d}t + \int_0^\eta \|\boldsymbol{r}_h\boldsymbol{h} - \boldsymbol{h}\|^2 \, \mathrm{d}t \le Ch^{2s},$$

and

$$\|(\boldsymbol{r}_h\boldsymbol{h}_0 - \boldsymbol{h}_0)\times\boldsymbol{n}\|^{\alpha}_{\boldsymbol{L}^{1+\alpha}(\Gamma)} + \|(\boldsymbol{r}_h\boldsymbol{h} - \boldsymbol{h})\times\boldsymbol{n}\|_{L^{1+\alpha}\left((0,T),\boldsymbol{L}^{1+\alpha}(\Gamma)\right)} \leq C\left(h^{\alpha(s-1/2)} + h^{s-1/2}\right).$$

□

## 4.7 Numerical experiments

The chapter concludes with performing two numerical experiments. We solve the test problem

$$\begin{aligned} \partial_t\boldsymbol{h} + \nabla\times\nabla\times\boldsymbol{h} + \boldsymbol{f} &= \boldsymbol{0} && \text{in} \quad \Omega\times(0,1), \\ \boldsymbol{n}\times(\nabla\times\boldsymbol{h}) &= \boldsymbol{n}\times\partial_t\boldsymbol{a}(\boldsymbol{h}\times\boldsymbol{n}) + \boldsymbol{n}\times\boldsymbol{g} && \text{on} \quad \Gamma\times(0,1), \\ \boldsymbol{h}(\boldsymbol{x},0) &= \boldsymbol{h}_0 && \text{in} \quad \Omega, \end{aligned} \tag{4.24}$$

where the domain $\Omega$ is the cube $(0,1)^3 \subset \mathbb{R}^3$ and the vector functions $\boldsymbol{f}$ and $\boldsymbol{g}$ will be defined later.

The numerical scheme follows the theoretical analysis. We discretize the problem in time according to the backward Euler scheme. On every time level, the Newton scheme is applied to deal with the nonlinearity. The space discretization is carried out by first order edge elements. We use the numerical software FEniCS [78].

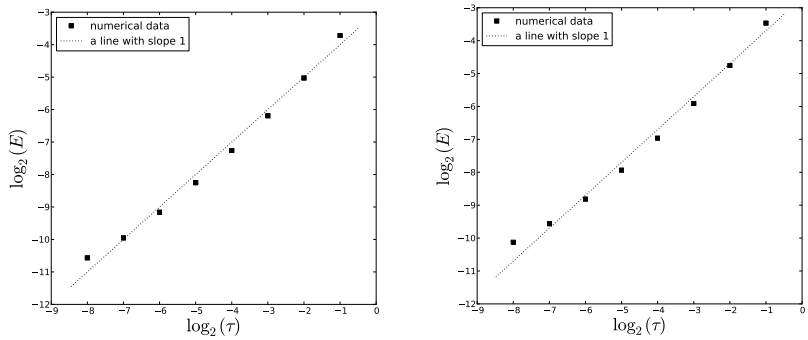
The first experiment is to test a convergence to an exact solution for the decreasing time step $\tau$. We choose $\boldsymbol{f}$ and $\boldsymbol{g}$ such that the vector field $\boldsymbol{h}(\boldsymbol{x},t) = (\sin(t)+1)(z-y, x-z, y-x)^\top$ represents the exact solution of (4.24). The vector field $\boldsymbol{h}(\boldsymbol{x},t)$ can be exactly fitted the first order edge elements and so a rather coarse mesh can be used. Figure 4.4 shows the absolute error $E$

$$E = \sqrt{\int_0^T \|\boldsymbol{h} - \boldsymbol{h}_{num}\|^2 \,\mathrm{d}t + \int_0^T \|\boldsymbol{h} - \boldsymbol{h}_{num}\|^2_\Gamma \,\mathrm{d}t}$$

between the numerical solution and the exact one in a log-log scale for the different times steps $\tau$. The numerical results have convergence rate $\mathcal{O}(\tau)$ what is optimal for first order methods. This is a better result than the convergence rate we proved in Theorem 4.3, what is probably caused by the smoothness of the solution.

Next, we investigate the dependence of the error on the discretization in both time and space. The test solution is $\boldsymbol{h}(\boldsymbol{x},t) = (t^2+1)(\sin(y)-\sin(z), \sin(z)-\sin(x), \sin(x)-\sin(y))^\top$. The fields $\boldsymbol{f}$ and $\boldsymbol{g}$ are defined in accordance with it. Figure 4.5 shows the absolute error for a fixed mesh and decreasing time step. With decreasing time step the space discretization error starts dominating over the time discretization error and so we see the flat part of the graph.

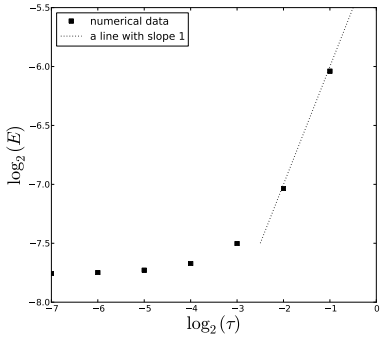
Figure 4.4: Absolute error with respect to the time step $\tau$



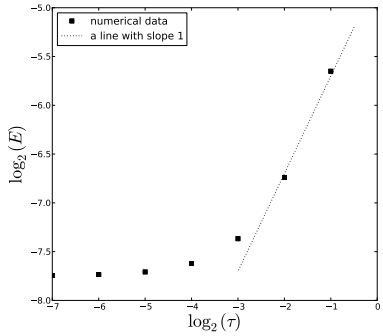

(a) $\alpha = 0.2$ (b) $\alpha = 0.5$

The behaviour of the space discretization error can be also inspected, if the time step $\tau$ is sufficiently small. Figure 4.6 shows the absolute error with respect to a refinement of mesh for the time step $\tau = 2^{-6}$. The space discretization parameter $h$ is given by Definition A.21.

Figure 4.5: Absolute error with respect to the time step $\tau$

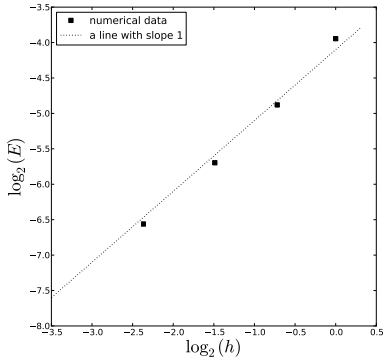



(a) $\alpha = 0.2$

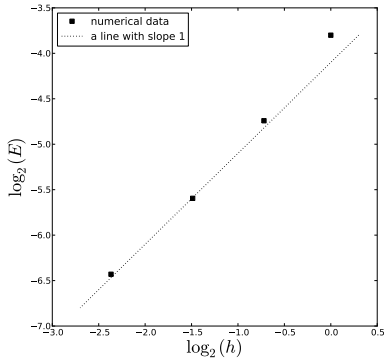



(b) $\alpha = 0.5$

Figure 4.6: Absolute error with respect to a mesh refinement ($\tau = 2^{-6}$)




(a) $\alpha = 0.2$




(b) $\alpha = 0.5$

# Part II

# On a continuation approach in Tikhonov regularization and its application in piecewise-constant parameter identification

# Chapter 5

# A continuation method for Tikhonov regularization

In this chapter we propose and study a *continuation-based approach* for Tikhonov regularization of *ill-posed* problems.

## 5.1 Introduction

We consider ill-posed problems that can be written in the form of an operator equation

$$Fu = v, \tag{5.1}$$

where $F : \mathcal{D}(F) \subseteq U \to V$ is a (in general non-linear) forward operator, mapping between Banach spaces $U$ and $V$. Its domain of definition is denoted by $\mathcal{D}(F)$ and its range by $\mathcal{R}(F)$. By $v$ we understand certain exact measurements projected on $V$. We assume that only noisy data $v^\delta$ are available, such that $\left\|v - v^\delta\right\|_V \leq \delta$, where $\delta$ is the level of noise.

Let us introduce a suitable *regularization* $\mathcal{R} : U \to [0, +\infty]$ with the domain $\mathcal{D}(\mathcal{R}) := \{u \in U : \mathcal{R}(u) \neq +\infty\}$. It is a proper and convex functional. The general convention is to consider only those solutions $u$ to the ill-posed operator equation (5.1), where $\mathcal{R}(u)$ is sufficiently small. An element $u^\dagger$ is called an $\mathcal{R}-$minimizing solution (e.g.[64]) if

$$\mathcal{R}(u^\dagger) = \min\{\mathcal{R}(u) : Fu = v\} < \infty. \tag{5.2}$$

We follow the classical Tikhonov idea [42, 85] and consider minimizers of the functional

$$\mathcal{T}_\alpha(u) := \left\|F(u) - v^\delta\right\|_V^2 + \alpha\mathcal{R}(u) \tag{5.3}$$

for a suitable regularization parameter $\alpha > 0$, which depends on both noise level and data, i.e. $\alpha = \alpha(\delta, v^\delta)$. The first term in (5.3) is called the *fidelity functional (term)*. It ensures that minima of the Tikhonov functional are approximate solutions of the operator equation (5.1), which is the problem we want to solve in the first place. The regularization term $\mathcal{R}(u)$ stabilizes the ill-posed problem with respect to the noise and represents a priori assumptions or expectations that we have about a desired solution. It practically always enforces the membership of $u$ in a certain $U$. As usual, we denote a minimizer of (5.3) as

$$u_\alpha^\delta := \operatorname*{argmin}_{u \in U} \mathcal{T}_\alpha(u). \tag{5.4}$$

It is a well known fact that under certain reasonable assumptions $u_\alpha^\delta$ are stable approximations of an $\mathcal{R}-$minimizing solution to (5.1), also in a rather general Banach space setting [64]. The resulting problem of regularization can be roughly stated as follows:

**Problem 5.1.** *Find a suitable $\alpha$ and the corresponding minimizer $u_\alpha^\delta$ of the Tikhonov functional (5.3), such that $u_\alpha^\delta$ approximates $u^\dagger$ as close as possible.*

Our main goal is to construct a sequence converging to the *global minimizer* $u_\alpha^\delta$. The biggest challenge is how to avoid the convergence of a numerical minimization method to a local minimum of (5.3), which is a common problem for standard *gradient-based minimization methods* (GBMM).

The possible reasons for the existence of local minima of (5.3) are triadic: the forward operator $F$ itself, the noise in the measurements and the penalty term $\mathcal{R}(u)$. The forward operator is case-specific and the noise is inherent to ill-posed problems. We have however full freedom of choice of regularization.

When a GBMM is applied to (5.3), the whole resulting minimizing sequence belongs to $U$. This is enforced by the regularization $\mathcal{R}(u)$. However the underlying direct problem (5.1) generally requires a far less regularity of a solution than it is asked by $\mathcal{R}(u)$. Even if we expect our final solution to belong to $U$, it is not necessary to consider only minimizing sequences from $U$. This restriction is often the reason that a GBMM converges to a local minimum.

Let us recall that the purpose of adding the regularization is to stabilize the ill-posed problem and to ensure the desired properties of the solution. The main idea will be to provide these two roles of the regularization term gradually.

### 5.1.1 Continuation immersion approach

Let us consider a Banach space $W$, such that $U$ is a proper subset of $W$ and the problem (5.1) is well defined in $W$, i.e. $U \subsetneq W$ and $\mathcal{D}(F) \cap W \neq \emptyset$. We can introduce a new Tikhonov functional analogical to (5.3)

$$\mathcal{T}_\beta(w) := \left\| F(w) - v^\delta \right\|^2 + \beta \mathcal{R}_W(w), \tag{5.5}$$

with a regularization term $\mathcal{R}_W : W \to [0, +\infty]$ and regularization parameter $\beta > 0$. It is again a convex and proper functional with the domain $\mathcal{D}(\mathcal{R}_W) := \{w \in W : \mathcal{R}_W(w) \neq +\infty\}$.

The main idea is to continuously transform the relaxed functional $\mathcal{T}_\beta$ to the original $\mathcal{T}_\alpha$ together with the corresponding minimization problems by making use of the continuation method [5]. We will stabilize the problem (5.1) using $W$-based regularization, i.e. in a "broader" sense. Since the Tikhonov regularization (5.5) in $W$ is a "less" constrained problem than (5.3), it will be easier solvable. It will provide a very good starting point for minimization in $U$. The extra desired properties will be progressively imposed on the solution via continuation-based projection a posteriori.

We consider a one-parameter family of the Tikhonov functionals

$$\mathcal{T}_{\alpha,\beta}(u, w, \lambda) = \left\|F(z) - v^\delta\right\|^2 + \lambda\alpha\mathcal{R}_U(u) + (1-\lambda)\beta\mathcal{R}_W(w), \tag{5.6}$$

where $\lambda \in [0, 1]$ and

$$z = \lambda u + (1-\lambda)w. \tag{5.7}$$

The regularization term $\mathcal{R}_U$ stands for the original regularization in (5.3). The regularizations parameters $\alpha$ and $\beta$ are in general functions of $\delta$, $v^\delta$. The forward operator $F$ corresponding to (5.6) can be understood as acting on the parametrized family $z \in W$. The regularization part

$$\mathcal{R}_{U,W}(u, w, \lambda) := \lambda\alpha\mathcal{R}_U(u) + (1-\lambda)\beta\mathcal{R}_W(w) \tag{5.8}$$

is better to be understood as a function on $U \times W$.

We consequently deal with a one-parameter family of minimization problems [1]. For $\lambda \in (0, 1)$ we look for a couple from $U \times W$, which minimizes the functional (5.6), that is

$$(u^\delta_{\alpha,\beta}(\lambda), w^\delta_{\alpha,\beta}(\lambda)) = \operatorname*{argmin}_{(u,w)\in U\times W} \mathcal{T}_{\alpha,\beta}(u, w, \lambda).$$

For $\lambda = 1$ we get the original minimization problem of $\mathcal{T}_\alpha$ and for $\lambda = 0$ the the problem reduces to the minimization of (5.5). By abuse of notation we sometimes write that $(u, w)$ is a minimizer of $\mathcal{T}_{\alpha,\beta}$ for any $\lambda \in [0, 1]$ to denote a minimizing couple $(u, w) \in U \times W$ if $\lambda \in (0, 1)$ and also to denote a minimizing element $w \in W$ if $\lambda = 0$ or $u \in U$ if $\lambda = 1$.

Analogically to the notion of the $R$-minimizing solution (5.2), let us define for each given $\lambda \in [0, 1]$ an $\mathcal{R}_{U,W}$-*minimizing solution* as a couple $(u^\dagger_\lambda, w^\dagger_\lambda) \in U \times W$, such that

$$\mathcal{R}_{U,W}(u^\dagger_\lambda, w^\dagger_\lambda, \lambda) = \min\{\mathcal{R}_{U,W}(u, w, \lambda) : F(z) = v\} < \infty.$$

[1] If $\alpha \equiv \beta$, the above formulas might bring the augmented Lagrangian method to mind. Among the differences between these two method, we stress that we minimize here in the two independent variables $u$ and $v$. This turns out very convenient, mainly from the numerical point of view, as we will show later.

## 5.2 Continuation approach for Tikhonov regularization

This section deals with theoretical aspects of the continuation approach for Tikhonov regularization. The functional $\mathcal{T}_{\alpha,\beta}$ defined by (5.6) is always minimized with respect to the variables $(u, w)$ and the variable $\lambda \in [0, 1]$ is taken as a fixed parameter

$$\mathcal{T}_{\alpha,\beta}(u, w, \lambda) \to \min, \quad \lambda u + (1 - \lambda)w = z \in \mathcal{D}(F). \tag{5.9}$$

Throughout the section we make the following assumptions:

(A1) Let $V$ be a Hilbert space and $W$ be a reflexive Banach space. The space $U$ is a closed reflexive proper subspace of $W$, $U \subsetneq W$.

(A2) $F : \mathcal{D}(F) \subseteq W \to V$, where $\mathcal{D}(F)$ is closed and convex, and $\mathcal{D} := \mathcal{D}(F) \cap U \neq \emptyset$. The map $F$ is a *strongly continuous*, i.e.

$$w_n \rightharpoonup w \quad \text{implies} \quad F(w_n) \to F(w). \tag{5.10}$$

It is furthermore a $C^1$-map.

(A3) $\mathcal{R}_W : W \to [0, \infty)$ is a $C^2$-map. It holds that $\mathcal{R}_W(0) = 0, \mathcal{R}'_W(0) = 0$ and the second derivative $\mathcal{R}''_W$ satisfies the condition

$$\langle \mathcal{R}''_W(w)h, h\rangle_{W^*} \geq C \|h\|^2_W$$

for any $w, h \in W$, where $C$ is a positive constant.

(A4) $\mathcal{R}_U : U \to [0, \infty)$ is a $C^2$-map. It holds that $\mathcal{R}_U(u) \geq \mathcal{R}_W(u)$ for any $u \in U$, $\mathcal{R}_U(0) = 0, \mathcal{R}'_U(0) = 0$ and the second derivative $\mathcal{R}_U$ satisfies the condition

$$\langle \mathcal{R}''_U(u)h, h\rangle_{U^*} \geq C \|h\|^2_U$$

for any $u, h \in U$, where $C$ is a positive constant.

Let us first make a few remarks about these assumptions. The assumption that $U \subsetneq W$ is general enough for our purposes. We will consider only Banach spaces $U$ that are continuously embeddable in a Lebesgue space $L^p := W$ for some $p \geq 1$. We assume that $V$ is a Hilbert space for simplicity. The results can be generalized for a Banach space $V$ with differentiable norm, see e.g. [115]. The following lemma will shed more light on the assumption (A2):

**Lemma 5.1.** *Let $F : \mathcal{D}(F) \subseteq W \to V$ be a strongly continuous operator between the reflexive Banach spaces $W$ and $V$. Then $F$ is completely continuous.*

*Proof.* All we need to prove is that $F$ is continuous and compact. The continuity of $F$ follows immediately from its strong continuity. Indeed, $w_n \to w$ implies $w_n \rightharpoonup w$ and so $F(w_n) \to F(w)$. To prove that $F$ is compact, it is sufficient to show that any bounded sequence $\{w_n\} \subset W$ is mapped onto the sequence $\{F(w_n)\} \subset V$ that contains strongly convergent subsequence. The reflexivity of $W$ implies that the bounded sequence $\{w_n\}$ contains a weakly convergent subsequence $\{w_m\}, w_m \rightharpoonup w$. Consequently, $\{F(w_m)\} \subset \{F(w_n)\}$ and $F(w_m) \to F(w) \in V$, which was to be proved. □

In the light of the above statement, we see that given the assumptions (A1) and (A2), the operator $F$ is completely continuous. This makes the problem (5.1) ill-posed as it has been shown in Section 1.2. We remark that the forward operator $F$ can be in many situations written as a composition of a map continuous in the weak topology and a compact embedding map, which is completely continuous map, e.g. Neumann-to-Dirichlet operators. The assumptions (A3) and (A4) imply that the regularizations $\mathcal{R}_U$ and $\mathcal{R}_W$ are convex proper functionals.

The first assertion provides a classical result about the existence of a minimizer and its characterization.

**Lemma 5.2** (Well-posedness)**.** *Assume (A1)-(A4). Let $\lambda \in [0,1]$ be arbitrary. Then there exists a minimizer of $\mathcal{T}_{\alpha,\beta}$ for any $\alpha \geq 0$, $\beta \geq 0$, which moreover satisfies the necessary condition*

$$\begin{aligned} D_u \mathcal{T}_{\alpha,\beta}(u, w, \lambda) &= 0, \\ D_w \mathcal{T}_{\alpha,\beta}(u, w, \lambda) &= 0. \end{aligned} \tag{5.11}$$

*If $\alpha$ and $\beta$ are large enough, then a critical point of $\mathcal{T}_{\alpha,\beta}$ is a local minimizer, i.e. the condition (5.11) is sufficient for a local minimum.*

*Proof.* The proof is a straightforward application of the variational calculus. Let $\lambda \in (0,1)$. We first show that $\mathcal{T}_{\alpha,\beta}$ is weakly sequentially lower semi-continuous. Since $F$ is strongly continuous and the norm is weakly sequentially lower semi-continuous, it holds that

$$\|F(z) - v\|_V^2 \leq \|F(z_n) - v\|_V^2 \quad \text{as} \quad w_n \rightharpoonup w \text{ and } u_n \rightharpoonup u.$$

The regularizations $\mathcal{R}_W$ and $\mathcal{R}_U$ are weakly sequentially lower semi-continuous by the continuity and convexity argument (Lemma A.2),

$$\mathcal{R}_W(w) \leq \mathcal{R}_W(w_n) \quad \text{as} \quad w_n \rightharpoonup w,$$

and

$$\mathcal{R}_U(u) \leq \mathcal{R}_U(u_n) \quad \text{as} \quad u_n \rightharpoonup u.$$

The functional $\mathcal{T}_{\alpha,\beta}$ is the conical sum of the above terms and hence it is weakly sequentially lower semi-continuous as well.

Now, Taylor's theorem shows for the regularization $\mathcal{R}_W$ that

$$\mathcal{R}_W(w) = \mathcal{R}_W(0) + \langle \mathcal{R}'_W(0), w\rangle_{W^*} + \int_0^1 (1-\theta)\, \langle \mathcal{R}''_W(\theta w)w, w\rangle_{W^*}\, \mathrm{d}\vartheta$$

and so from the assumption (A3) we conclude

$$\mathcal{R}_W(w) \geq C\, \|w\|_W^2 \qquad \text{for any } w \in W. \tag{5.12}$$

Analogously, it follows from the assumption (A4) that

$$\mathcal{R}_U(u) \geq C\, \|u\|_U^2 \qquad \text{for any } u \in U. \tag{5.13}$$

This shows that the functional $\mathcal{T}_{\alpha,\beta}$ is also weakly coercive, i.e.

$$\mathcal{T}_{\alpha,\beta}(u, w, \lambda) > C\left(\lambda\alpha\, \|u\|_U^2 + (1-\lambda)\beta\, \|w\|_W^2\right) \to \infty \quad \text{ as } \|u\|_U + \|w\|_W \to \infty.$$

The weak sequential lower semi-continuity and coercivity imply that the functional $\mathcal{T}_{\alpha,\beta}$ attains its minimum (see Theorem A.13 ). As $\mathcal{T}_{\alpha,\beta}$ is Gâteaux differentiable, a minimizer solves the equation (5.11). The case when $\lambda = 0$ and $\lambda = 1$ follows the same lines. One has to consider only $w$ and $u$ variable, respectively.

The second derivative of $\mathcal{T}_{\alpha,\beta}$ with respect to $u$ and $w$ is positive for some sufficiently large $\alpha$ and $\beta$, which implies that every solution of (5.11) is a local minimizer (see Theorem A.14). □

Expanding the condition (5.11) for $\lambda \in (0,1)$ reveals [2]

$$\Big[2\,(F'(z)\,\cdot, F(z) - v) + \alpha\, \langle \mathcal{R}'_U(u), \cdot\rangle_{U^*}\Big]\lambda = 0,$$
$$\Big[2\,(F'(z)\,\cdot, F(z) - v) + \beta\, \langle \mathcal{R}'_W(w), \cdot\rangle_{W^*}\Big](1-\lambda) = 0,$$

and thus

$$\alpha\, \langle \mathcal{R}'_U(u), \cdot\rangle_{U^*} = \beta\, \langle \mathcal{R}'_W(w), \cdot\rangle_{W^*}.$$

We use the above formula to establish the so-called Ritz projection from the space $W$ to its subspace $U$, which will turn out useful.

**Lemma 5.3** (Ritz projection)**.** *Assume (A1), (A3) and (A4). Let $u \in U$ be the solution of the problem*

$$\alpha\, \langle \mathcal{R}'_U(u), h\rangle_{U^*} = \beta\, \langle \mathcal{R}'_W(w), h\rangle_{W^*} \qquad \textit{for all } h \in U, \tag{5.14}$$

*where $w \in W$ and $\alpha, \beta > 0$ are fixed. Then,*

[2] Note that $F' : W \to L(W, V)$, and so $F'(z) \in L(W, V)$ for $z \in W$ and $F'(z)h \in V$ for $h \in W$

*(i) the map $\mathcal{P} : W \to U$ such that $w \mapsto \mathcal{P}(w) = u$ is well-defined,*

*(ii) the map $\mathcal{P}$ is continuously differentiable with $\mathcal{P}' = [\mathcal{R}''_U(\mathcal{P}(w))]^{-1} \circ \frac{\beta}{\alpha}\mathcal{R}''_W$,*

*(iii) the a priori estimate $\|u\|_U \leq C \|\mathcal{R}'_W(w)\|_{L(W,W^*)}$ holds true.*

*Proof.* (i) It is sufficient to prove the unique solvability of the problem (5.14). Since $U \subset W$, it follows that $W^* \subset U^*$, and hence $\mathcal{R}'_W(w) \in U^*$. The assumption (A4) implies that the operator $\mathcal{R}'_U : U \to U^*$ is hemicontinuous, i.e. the map $t \mapsto \langle \mathcal{R}'_U(u_1 + tu_2), h\rangle_{U^*}$ is continuous on $[0,1]$ for all $u_1, u_2, h \in U$. We furthermore deduce that

$$\begin{aligned}
&\langle \mathcal{R}'_U(u_1) - \mathcal{R}'_U(u_2), u_1 - u_2\rangle_{U^*} \\
&= \left\langle \int_0^1 \mathcal{R}''_U(u_1 + \theta(u_2 - u_1))(u_1 - u_2)\,\mathrm{d}\vartheta, u_1 - u_2 \right\rangle_{U^*} \\
&= \int_0^1 \langle \mathcal{R}''_U(u_1 + \theta(u_2 - u_1))(u_1 - u_2), u_1 - u_2\rangle_{U^*}\,\mathrm{d}\vartheta \\
&\geq C \|u_1 - u_2\|_U^2 ,
\end{aligned}$$

which shows that $\mathcal{R}'_U$ is strongly monotone and a fortiori coercive. The theory of monotone operators (see Theorem A.10) then guarantees that for any $w \in W$ there exists a unique $u = \mathcal{P}(w)$ such that

$$\alpha \mathcal{R}'_U(\mathcal{P}(w)) = \beta \mathcal{R}'_W(w), \tag{5.15}$$

and that $[\mathcal{R}'_U(u)]^{-1}$ is Lipschitz continuous.

(ii) We can now apply the local inverse function theorem (Theorem A.8), because the derivative $\mathcal{R}''_U(\mathcal{P}(w)) \in L(U, U^*)$ is bijective on account of (A4) and the linear operator theory. It is furthermore a global inverse map, because $\mathcal{R}'_U$ is proper, i.e. the preimage $\mathcal{R}'_U(M)$ of any compact set $M$ is also compact (Theorem A.9). Consequently, the differentiation of (5.15) yields

$$\mathcal{P}'(w) = \left[\mathcal{R}''_U(\mathcal{P}(w))\right]^{-1} \circ \frac{\beta}{\alpha}\mathcal{R}''_W(w), \qquad w \in W.$$

(iii) We put $h = u$ in (5.14) to estimate that

$$\begin{aligned}
C \|u\|_U^2 &\leq \alpha \langle \mathcal{R}'_U(u), u\rangle_{U^*} = \beta \langle \mathcal{R}'_W(w), u\rangle_{W^*} \\
&\leq \beta \|\mathcal{R}'_W(w)\|_{L(W,W^*)} \tilde{C} \|u\|_U ,
\end{aligned}$$

which concludes the proof. □

**Remark 5.1.** *Let us give a simple example. Suppose that* $\alpha = \beta$, $\mathcal{R}_W = \|\cdot\|_W^2$ *and* $\mathcal{R}_U = \|\cdot\|_W^2 + \mathcal{S}(u)$ *where* $\mathcal{S}$ *is a seminorm on* $U$. *The problem (5.14) is then equivalent to the minimization problem*

$$\min_{u \in U} \left[ \|u - w\|_W^2 + \mathcal{S}(u) \right].$$

*Thus the projection* $\mathcal{P}$ *looks for an element* $u \in U$ *which is the closest one to a* $w \in W$ *and in the same time has the minimal seminorm* $\mathcal{S}$.

**Remark 5.2.** *The direct consequence of the above considerations is that the system (5.11) is for* $\lambda \in (0, 1)$ *equivalent to the system*

$$\begin{aligned} D_w \mathcal{T}_{\alpha,\beta}(\mathcal{P}(w), w, \lambda) &= 0, \\ \alpha \mathcal{R}'_U(\mathcal{P}(w)) &= \beta \mathcal{R}'_W(w), \end{aligned}$$

*and for* $\lambda = 0$ *we can still define "the minimizer"* $u^\delta_{\alpha,\beta}(0)$ *as the projection* $\mathcal{P}(w^\delta_{\alpha,\beta}(0))$.

The following theorem provides the main result of this section. It establishes a continuous dependence of the minimizer of $\mathcal{T}_{\alpha,\beta}$ on the parameter $\lambda$. The main idea of the proof lies in realizing that the problem is a saddle point one. We minimize in $U \times W$ and maximize in $\lambda$. Further, the proof follows the standard lines (compare with [42]).

**Theorem 5.1** (Continuous dependence on $\lambda$)**.** *Assume (A1)-(A4). Let* $\alpha \geq \beta > 0$ *and* $v^\delta \in V$. *Assume that there exists a unique global minimizer* $(u^\delta_{\alpha,\beta}(\lambda), w^\delta_{\alpha,\beta}(\lambda))$ *of (5.6) for any* $\lambda \in [0, 1]$ [3]. *Then the mappings*

$$\begin{aligned} w^\delta_{\alpha,\beta} : [0, 1) \to W, \qquad & \lambda \mapsto w^\delta_{\alpha,\beta}(\lambda), \\ u^\delta_{\alpha,\beta} : (0, 1] \to U, \qquad & \lambda \mapsto u^\delta_{\alpha,\beta}(\lambda) \end{aligned}$$

*are continuous.*

The theorem has an important corollary, which establishes a local correctness of the continuation extension at $\lambda = 1$ :

**Corollary 5.1.** *Let the assumptions of Theorem 5.1 be fulfilled. If* $\lambda \to 1$, *then* $u^\delta_{\alpha,\beta}(\lambda) \to u^\delta_\alpha$.

*Proof.* We begin the proof of Theorem 5.1 with a few estimates for $\mathcal{T}_{\alpha,\beta}$, which will help us later. It is evident that

$$\mathcal{R}_{U,W}(u, v, \lambda) \leq \alpha \mathcal{R}_U(u) + \beta \mathcal{R}_W(w) \tag{5.16}$$

[3] As we have mentioned, if $\lambda = 0$ and $\lambda = 1$, we consider just $w^\delta_{\alpha,\beta}(0)$ and $u^\delta_{\alpha,\beta}(1)$, respectively.

for any $u \in U, w \in W$ and $\lambda \in [0,1]$ . Conversely, the assumption (A4) and the convexity of $\mathcal{R}_W$ imply

$$\begin{aligned}\mathcal{R}_{U,W}(u,w,\lambda) &\geq \alpha\lambda\mathcal{R}_W(u) + \beta(1-\lambda)\mathcal{R}_W(w)\\ &\geq \beta[\lambda\mathcal{R}_W(u) + (1-\lambda)\mathcal{R}_W(w)]\\ &\geq \beta\mathcal{R}_W(\lambda u + (1-\lambda)w),\end{aligned}$$

which leads to the estimate

$$\left\|F(z) - v^\delta\right\|_V^2 + \beta\mathcal{R}_W(z) \leq \left\|F(z) - v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u,w,\lambda) \tag{5.17}$$

for any $(u,w) \in \mathcal{D} \times \mathcal{D}(F)$ and $z = \lambda u + (1-\lambda)w$. By the mean value theorem we obtain for the fidelity term

$$\begin{aligned}\left\|F(z) - v^\delta\right\|_V^2 &= \left\|F(z) \mp F(w) - v^\delta\right\|_V^2\\ &\leq 2\left(\|F(z) - F(w)\|_V^2 + \left\|F(w) - v^\delta\right\|_V^2\right)\\ &\leq 2\left(\|F'(\xi)(\lambda u + (1-\lambda)w - w)\|_V^2 + \left\|F(w) - v^\delta\right\|_V^2\right)\\ &\leq 2\left(\left[\|F'\|_{L(S,V)}\,\lambda\,\|u-w\|_W\right]^2 + \left\|F(w) - v^\delta\right\|_V^2\right),\end{aligned} \tag{5.18}$$

where the set $S$ is the line segment $u + t(w-u) \in W$, $t \in [0,1]$.

Let now $\lambda_k \to \lambda \in [0,1]$ as $k \to \infty$. Denote by $(u_k, w_k)$ the corresponding global minimizer $(u^\delta_{\alpha,\beta}(\lambda_k), w^\delta_{\alpha,\beta}(\lambda_k))$ and set $z_k = \lambda_k u_k + (1-\lambda_k)w_k$. By the definition of minimizer it holds true that

$$\mathcal{T}_{\alpha,\beta}(u_k, w_k, \lambda_k) \leq \mathcal{T}_{\alpha,\beta}(u, w, \lambda_k)$$

for any $(u,w) \in \mathcal{D} \times \mathcal{D}(F)$. We can moreover bound the minimum of $\mathcal{T}_{\alpha,\beta}$ uniformly for any $\lambda \in [0,1]$ with the estimates (5.16) and (5.18)

$$\begin{aligned}&\left\|F(z_k) - v^\delta\right\|_V^2 + \alpha\lambda_k\mathcal{R}_U(u_k) + \beta(1-\lambda_k)\mathcal{R}_W(w_k)\\ \leq&\left\|F(z) - v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u,w,\lambda_k)\\ \leq&2\left[\|F'\|_{L(S,V)}\,\|u-w\|_W\right]^2 + 2\left\|F(w) - v^\delta\right\|_V^2 + \alpha\mathcal{R}_U(u) + \beta\mathcal{R}_W(w),\end{aligned} \tag{5.19}$$

where $(u,w) \in \mathcal{D} \times \mathcal{D}(F)$. This implies combining with (5.12) and (5.13) that

$$C(1-\lambda_k)\,\|w_k\|_W^2 \leq \beta(1-\lambda_k)\mathcal{R}_W(w_k) \leq \tilde{C},$$

and

$$C\lambda_k\,\|u_k\|_U^2 \leq \alpha\lambda_k\mathcal{R}_U(u_k) \leq \tilde{C}.$$

Therefore, the sequences $\{u_k\}$ and $\{w_k\}$ are bounded in $U$ and $W$, unless $\lambda_k \to 0$ and $\lambda_k \to 1$, where the estimate (5.19) is unapplicable for $\{u_k\}$ and $\{w_k\}$, respectively. If $\lambda_k \to 0$, we can however use Lemma 5.3 to find

$$\|u_k\|_U \leq \mathcal{R}'_W(w_k) \leq C.$$

and consequently

$$\lambda_k u_k \to 0 \quad \text{in } U \quad \text{as } \lambda_k \to 0.$$

If $\lambda_k \to 1$, it follows from

$$C(1-\lambda_k)\|w_k\|_W^2 = C\left\|\sqrt{1-\lambda_k}w_k\right\|_W^2 \leq \tilde{C}$$

that

$$(1-\lambda_k)w_k \to 0 \quad \text{in } W \quad \text{as } \lambda_k \to 1.$$

The estimates (5.17) and (5.12) on the other hand force

$$\begin{aligned}\left\|F(z_k)-v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u_k,w_k,\lambda_k) &\geq \left\|F(z_k)-v^\delta\right\|_V^2 + \beta\mathcal{R}_W(z_k)\\ &\geq \beta C\|z_k\|_W^2,\end{aligned} \tag{5.20}$$

which together with (5.19) ensures that the sequence $\{z_k\}$ is always uniformly bounded in $W$

$$\|z_k\|_W \leq C.$$

Bounded sequences in reflexive spaces are weakly compact and so we can choose weakly convergent subsequences

$$u_m \rightharpoonup \overline{u},\ w_m \rightharpoonup \overline{w} \text{ and } z_m \rightharpoonup \overline{z} \quad \text{as } m \to \infty. \tag{5.21}$$

The above estimates moreover establish that

$$\overline{z} = \lambda\overline{u} + (1-\lambda)\overline{w} \quad \text{for any } \lambda \in [0,1].$$

We then consecutively deduce by the weak sequential lower semi-continuity of $\mathcal{T}_{\alpha,\beta}$ and the definition of minimizer that

$$\begin{aligned}&\|F(\overline{z})-v^\delta\|_V^2 + \mathcal{R}_{U,W}(\overline{u},\overline{w},\lambda)\\ &\quad\leq \liminf_{m\to\infty}\left[\left\|F(z_m)-v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u_m,w_m,\lambda_m)\right]\\ &\quad\leq \limsup_{m\to\infty}\left[\left\|F(\lambda_m u_m+(1-\lambda_m)w_m)-v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u_m,w_m,\lambda_m)\right]\\ &\quad\leq \lim_{m\to\infty}\left[\left\|F(\lambda_m u+(1-\lambda_m)w)-v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u,w,\lambda_m)\right]\\ &\quad= \left\|F(z)-v^\delta\right\|_V^2 + \mathcal{R}_{U,W}(u,w,\lambda)\end{aligned}$$

for all $(u, w) \in \mathcal{D} \times \mathcal{D}(F)$. This shows that $(\overline{u}, \overline{w})$ is minimizer of (5.9) and that

$$\lim_{m\to\infty} \mathcal{T}_{\alpha,\beta}(u_m, w_m, \lambda_m) = \mathcal{T}_{\alpha,\beta}(\overline{u}, \overline{w}, \lambda). \tag{5.22}$$

Assume now that $(u_m, w_m)$ does not strongly converge to $(\overline{u}, \overline{w})$. Then, the sequence $\{(u_m, w_m)\}$ is still bounded and so there exists a subsequence $\{(u_n, w_n)\}$ of $\{(u_m, w_m)\}$ such that $(u_n, w_n) \rightharpoonup (\overline{u}, \overline{w})$, $F(z_n) \rightharpoonup F(\overline{z})$ and $\mathcal{R}_{U,W}(u_n, w_n, \lambda) \to c$, where

$$c := \limsup \mathcal{R}_{U,W}(u_m, w_m, \lambda) > \mathcal{R}_{U,W}(\overline{u}, \overline{w}, \lambda).$$

As a consequence of (5.22), we obtain

$$\begin{aligned}\lim_{n\to\infty} \left\|F(z_n) - v^\delta\right\|_V &= \left\|F(\overline{z}) - v^\delta\right\|_V + \mathcal{R}_{U,W}(\overline{u}, \overline{w}, \lambda) - c \\ &< \left\|F(\overline{z}) - v^\delta\right\|_V,\end{aligned}$$

which is in contradiction with the weak lower semicontinuity of the norm.

Since the minimizer $(\overline{u}, \overline{w})$ is unique for any $\lambda \in [0, 1]$, the above considerations demonstrate that every sequence $\{(u_k, w_k)\}$ contains a subsequence strongly converging towards $(\overline{u}, \overline{w})$, and therefore, the functions $u^\delta_{\alpha,\beta}$ and $w^\delta_{\alpha,\beta}$ are continuous on the intervals $(0, 1]$ and $[0, 1)$, respectively. □

The next two theorems address the questions of stability and convergence of minimizers of $\mathcal{T}_{\alpha,\beta}$. We omit their proofs, because they go along the same lines as in e.g. [42, Theorem 10.2 and 10.3].

**Theorem 5.2** (Stability)**.** *Assume (A1)-(A4), $\alpha > 0, \beta > 0$ and $v^\delta \in V$. Let $\lambda \in [0, 1]$ be fixed and let $\{v_k\}$ and $\{(u_k, w_k)\}$ be sequences such that $v_k \to v^\delta$ and $(u_k, w_k)$ is a minimizer of (5.6) with $v^\delta$ replaced by $v_k$. Then there exists a convergent subsequence of $\{(u_k, w_k)\}$ and the limit of every convergent subsequence is a minimizer of (5.6).*

**Theorem 5.3** (Convergence)**.** *Assume (A1)-(A4). Let $v^\delta \in V$ with $\left\|v - v^\delta\right\|_V \le \delta$ and let $\lambda \in [0, 1]$ be fixed. Let $\alpha(\delta)$ and $\beta(\delta)$ be such that $\alpha(\delta) \to 0$, $\beta(\delta) \to 0$ and $\delta^2/\alpha(\delta) \to 0$, $\delta^2/\beta(\delta) \to 0$ as $\delta \to 0$. Then every sequence $\{(u^{\delta_k}_{\alpha_k}, w^{\delta_k}_{\beta_k})\}$, where $\delta_k \to 0$, $\alpha_k = \alpha(\delta_k)$, $\beta_k = \beta(\delta)$ and $(u^{\delta_k}_{\alpha_k}, w^{\delta_k}_{\beta_k})$ is the solution of (5.9), has a convergent subsequence. The limit of every convergent subsequence is an $\mathcal{R}_{U,W}$-minimizing solution. If in addition, the $\mathcal{R}_{U,W}$-minimizing solution $(u^\dagger_\lambda, w^\dagger_\lambda)$ is unique, then*

$$\lim_{\delta\to 0}(u^{\delta_k}_{\alpha_k}, w^{\delta_k}_{\beta_k}) = (u^\dagger_\lambda, w^\dagger_\lambda).$$

The last result about the existence of an $\mathcal{R}_{U,W}$-minimizing solution is essentially due to [64].

**Lemma 5.4.** *Assume (A1)-(A4). If there exists a solution of (5.1), then there exists an $\mathcal{R}_{U,W}$-minimizing solution for any $\lambda \in [0, 1]$.*

*Proof.* Let $v^\delta = v$ in (5.6) and consider the case when $\lambda \in (0,1)$. Suppose for the sake of contradiction that there does not exist an $\mathcal{R}_{U,W}$-minimizing solution in $\mathcal{D} \times \mathcal{D}(F)$. Then there exists a sequence $\{(u_k, w_k)\}$ of solutions of (5.1) in $\mathcal{D} \times \mathcal{D}(F)$ such that $\mathcal{R}_{U,W}(u_k, w_k, \lambda) \to c$ and

$$\begin{aligned} c &< \mathcal{R}_{U,W}(u, w, \lambda) \\ \text{for all } (u,w) \in U \times V \text{ satisfying} \quad & F(\lambda u + (1-\lambda)w) = v. \end{aligned} \tag{5.23}$$

For a sufficiently large $k$, it follows that

$$\mathcal{T}_{\alpha,\beta}(u_k, w_k, \lambda) = \mathcal{R}_{U,W}(u, w, \lambda) < 2c,$$

and so we see by (5.13) and (5.12) that

$$C\left(\lambda\alpha \left\|u_k\right\|_U^2 + (1-\lambda)\beta \left\|w_k\right\|_W^2\right) \leq 2c. \tag{5.24}$$

One can thus extract a weakly convergent subsequence, again denoted by $\{(u_k, w_k)\}$, with the limit $(\overline{u}, \overline{w})$. The weak lower semicontinuity of $\mathcal{R}_{U,W}$ implies that

$$\mathcal{R}_{U,W}(\overline{u}, \overline{w}, \lambda) \leq \liminf_{k\to\infty} \mathcal{R}_{U,W}(u_k, w_k, \lambda) = c.$$

However, the map $F$ is strongly continuous and hence the equality $F(\lambda u_k + (1-\lambda)w_k) = v$ forces $F(\lambda\overline{u} + (1-\lambda)\overline{w}) = v$, which is the contradiction to (5.23).

The case when $\lambda = 0$ and $\lambda = 1$ goes along the same lines. One has to consider only $\mathcal{R}_W$ and $\mathcal{R}_U$ functionals with corresponding $\mathcal{R}_W$-minimizing solution and $\mathcal{R}_U$-minimizing solution, respectively.

□

# Chapter 6

# Piecewise-constant parameter identification

This chapter discusses the above presented continuation approach for Tikhonov regularization in context of piecewise-constant parameter identification problems. Subsequently, we present its application to magnetic induction tomography.

## 6.1 Piecewise-constant parameter identification problems

Our motivation to study minimizers of (5.6) comes from piecewise-constant parameter identification problems (PIPs). We analyze partial differential equation (PDE) constrained problems with the unknown parameter being a coefficient of the PDE-constraint.

Let us consider a double-valued piecewise-constant parameter $\sigma$. This case is general enough and suitable for illustration purposes and we will hold this assumption throughout this chapter. We have

$$\sigma_{PC} = \sigma_1 \chi_D + \sigma_2 \chi_{\Omega/D}, \quad \sigma_1, \sigma_2 \in \mathbb{R}, \tag{6.1}$$

where the domain $\Omega$ is an open bounded set, on which the PDE-constrained problem is defined. The symbols $\chi_D$ and $\chi_{\Omega/D}$ stand for the characteristic function of subset $D \subset \Omega$ and its complement, respectively. The goal is to find the subdomain $D$ and the unknown numbers $\sigma_1$ and $\sigma_2$ based on suitable observations of the state variable of the PDE-constraint.

A classical example here is the problem of inverse electric impedance tomography (EIT) [17]. The aim of EIT is to determine the unknown conductivity $\sigma$ from the following Neumann-to-Dirichlet operator

$$\Lambda_\sigma : g \mapsto u|_\Gamma, \tag{6.2}$$

which is well-defined by the unique solution of the governing equation for the electric potential $u$

$$\begin{aligned} \nabla \cdot (\sigma \nabla u) &= 0 \quad \text{in } \Omega, \\ \sigma \nabla u \cdot \boldsymbol{n} &= g \quad \text{on } \Gamma. \end{aligned}$$

The domain $\Omega \subset \mathbb{R}^2$ represents the imaged body, $\Gamma = \partial\Omega$ its surface and $g = g(x)$ an applied current. An exact derivation of this model from the Maxwell equations can be found for instance in [73]. Applying the divergence theorem to (6.1) leads to the conservation of charge condition $\int_\Gamma g \, \mathrm{d}S = 0$, which is a necessary condition for the existence of a solution $u$. Given the normative condition $\int_\Gamma u \, \mathrm{d}S = 0$, there exists the unique solution $u$ for any positive and bounded conductivity $\sigma$

$$0 < \sigma_{\min} < \sigma < \sigma_{\max}.$$

We include EIT into our considerations simply by putting $F := \Lambda_\sigma$ and considering only the piecewise constant $\sigma = \sigma_{PC}$ as defined in (6.1).

We are primary concerned by building an efficient numerical algorithm to recover the unknown $\sigma$ for problems like EIT. In the case of EIT, the problematic is extensively studied in the literature. We again refer to [17] for a good review of both non-iterative and iterative methods for EIT. The electric induction tomography is an severely ill-posed inverse problem. A crucial question of uniqueness in $L^\infty$ has been positively answered in the relatively recent paper [10]. The paper provides a final answer to the famous Caldern's problem (see the reprint [24] of the original article from 1980) which has been attracting a lot of attention [110, 66].

Why is it reasonable to look for the optimal $\sigma$ in the space of piecewise constant functions? Such a choice is natural, given a problem like EIT. First, this class of functions is rich enough in order to be applicable. Second, as in the case of EIT, one usually has only a finite number of measurements on the boundary $\Gamma$ corresponding to the Neumann-to-Dirichtet operator. For a two-dimensional domain $\Omega$, these measurements are one-dimensional. It is reasonable to assume that we can successfully recover at most a one-dimensional unknown inside the domain. [1] It is precisely, what one does by considering (6.1). The goal is as a matter of fact to find the interface between the two regions of $\Omega$. It is the choice of space plays a role of regularization.

$\mathbf{U} = \mathbf{BV}(\Omega)$: The most suitable type of regularization for piecewise-constant parameter identification problems is the $BV(\Omega)-$ regularization [1]. The space $BV(\Omega)$ is the subspace of functions $u \in L^1(\Omega)$ such that the quantity

$$J(u, \Omega) = \sup \left\{ \int_\Omega u(x) \nabla \cdot \xi(x) \, \mathrm{d}x : \xi \in C_c^\infty(\Omega, \mathbb{R}^n), \|\xi\|_{L^\infty(\Omega, \mathbb{R}^n)} \leq 1 \right\},$$

is finite, where $C_c^\infty(\Omega, \mathbb{R}^n)$ is the set of smooth functions in $C^\infty(\mathbb{R}^n)$ with compact

[1] We do not claim that certain two-dimensional recovery is impossible.

support in $\Omega$. Endowed with the norm

$$\|u\|_{BV(\Omega)} := \|u\|_{L^1(\Omega)} + J(u, \Omega), \tag{6.3}$$

it is a Banach space.

Tikhonov regularization formulation for the piecewise-constant PIP then reads as

$$\mathcal{T}_\alpha(\sigma_{PC}) := \left\|F(\sigma_{PC}) - v^\delta\right\|_V^2 + \alpha \left\|\sigma_{PC}\right\|_{BV(\Omega)}^2, \tag{6.4}$$

where $F$ is the operator associated with the forward problem, e.g. $\Lambda_\sigma$ for $V = L^2(\Gamma)$. This functional is a particular case of the functional (5.3) from the introduction when we set $U = BV(\Omega)$.

### 6.1.1 State of the art of geometry (shape) identification

In the case that the constants $\sigma_1$ and $\sigma_2$ in (6.1) are identified, the piecewise-constant parameter $\sigma$ estimation is equivalent to the geometry identification of the subdomain $D$.

The classical methods to identify the structural information are mostly based on a study of the sensitivity of a certain cost functional to a infinitesimal change of the shape of the structure itself, see [90] and the references therein. This shape sensitivity approach yields eventually to the notion of *shape derivative* [107].

The methods based on the shape sensitivity approach, level set method parameterizations including [97, 47], are updating the shape of the domain first, not the topology. The topology is prescribed a priori by an initial guess. The choice of a good initial guess becomes very important for the method to converge to the optimal shape. Even if some proposed (and well designed) algorithms are able to find the optimal shape [28], the convergence is usually very slow. The speed of the convergence is again strongly dependent on the good initial guess.

The second class of methods is based on the *homogenization* theory, see the pioneering work [15] or the monograph [2]. The optimal geometry is obtained in an enriched space of composite designs. For example, in a two phase optimization problem one looks for an optimal distribution of two components which minimizes a suitable objective function. If the conductivity of one component is allowed to go to zero, then, in the limit, this component models the voids. The final composite design is described by the material density function. The corresponding classical design can be retrieved via thresholding or penalization.

This approach overcomes some restrictions of the classical shape sensitivity approach. Both the topology and shape are optimized at once. The final acquired geometries are close to the optimal onces. Unfortunately, this approach is limited to certain types of problems and its rigorous application is a non-trivial task.

In this context we would like to mention the use of convexification or the so-called *fictitious material approach* for structural optimization (in this case the homogenization

approach is equivalent to the quasiconvexification), see the pioneering work [26]. The method is less theoretically sounded than the homogenization approach and the obtained optimal designs are inferior to that of the homogenization approach as well [2]. But the idea is straightforward to implement to almost any problem in mind.

A method based on an iterative inclusion of new holes (so called "bubbles") into the geometry was investigated in [44]. This idea is actually closely related to the one of the homogenization approach. In [98], a pointwise limit of such inclusions was used in linear elasticity to find a optimal design characterized by the so-called compliance functional. The importance of this contribution was recognized in [104, 105, 106], where the notion of *topological derivative* was introduced and further developed. Assume that $\Omega \subset \mathbb{R}^n$ is an open set and that there is given a shape functional

$$\mathcal{E} : \Omega/K \to \mathbb{R} \tag{6.5}$$

for any compact subset $K \subset \Omega$. We denote by $B_r(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, the ball of radius $r > 0$, i.e. $B_r(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n : |\boldsymbol{y} - \boldsymbol{x}| < r\}$. $\overline{B_r(\boldsymbol{x})}$ is the closure of $B_r(\boldsymbol{x})$. Assume that there exists the following limit

$$\mathcal{T}(\boldsymbol{x}) = \lim_{r \to 0} \frac{\mathcal{E}(\Omega \setminus \overline{B_r(\boldsymbol{x})}) - \mathcal{E}(\Omega)}{|\overline{B_r(\boldsymbol{x})}|}. \tag{6.6}$$

The function $\mathcal{T}(\boldsymbol{x}), \boldsymbol{x} \in \Omega$ is called the topological derivative of $\mathcal{E}(\Omega)$ and provides the information about the infinitesimal variation of the shape functional $\mathcal{E}$ if a small hole is created at $\boldsymbol{x} \in \Omega$.

Since the introduction of the topological derivative, a great number of contributions were made using this concept both in science and in engineering. We are interested particularly in those where topological and shape sensitivity concepts are used in conjunction.

In [23] the authors first considered the shape derivative based level set method (LSM). The motion of the interface described by the LSM is governed by a non-linear Hamilton-Jacobi equation, where its speed is dependent on the shape derivative of the cost functional. The idea was to introduce a new source term into the Hamilton-Jacobi equation, dependent on the topological derivative. This term allows for nucleation of new holes in the domain. The approach was generalized in [60].

In [4] the authors study the shape derivative based level set method for structural optimization. They do not use the topological derivative in the work itself, but, to our best knowledge, for the first time the topological derivative is suggested to be used for initialization of the algorithms based on the shape sensitivity approach. They study the idea in [3], where an alternating algorithm using both the shape and the topological derivatives is proposed.

In [87] the authors propose a variant of a binary level set approach for solving elliptic problems with piecewise constant coefficients. The inverse problem is solved by a

variational augmented Lagrangian approach with a total variation regularization. Their implementation was able to recover rather complicated geometries without assuming anything about $D$ a priori, i.e. without any initial guess. As we will understand later on, it is due to the nature of the augmented Lagrangian approach, which imposes the piecewise constant constraint gradually. The results of [87] are applied to the piecewise constant level set method (PCLSM) parametrization in [128]. They are employed to study an optimization problem. The PCLSM methods for the identification of discontinuous parameters in ill-posed problems are considered in [37]. Both a Tikhonov regularization approach using operator splitting techniques and an augmented Lagrangian approach are introduced and analyzed.

In [62] topological sensitivity based initial guess is used as a starting point for the shape-sensitivity level set method to solve an electric impedance tomography problem.

### 6.1.2 Topology-to-shape continuation method

In this section we introduce a continuation approach to shape identification, which combines topology and shape sensitivities.

The main idea is based on the following reasoning. Roughly speaking, topological properties of a particular shape are those which stay invariant under various *continuous transformations* [2]. A shape itself is a certain topology modified by those continuous boundary-like transformations, see the above section. Therefore, the topology is the "coarse" information about a particular shape. In this line of reasoning, it is intuitive to first look for the topology itself and to consider *continuation* methods to transform it to the particular shape.

We will consider the relaxed parametrization of $\sigma_{PC}$

$$\sigma = (1-\lambda)\sigma_{L^2} + \lambda\sigma_{PC}, \tag{6.7}$$

analogously to (5.7). We assume that $\sigma_{L^2} \in L^2(\Omega)$, because the space $U = BV(\Omega)$ is included at most in $W = L^2(\Omega)$ in the case if the domain $\Omega \subset \mathbb{R}^2$.

The function $\sigma_{L^2}$ can be interpreted as the topological derivative. It is almost everywhere locally defined and represents the distribution of the mass in $\Omega$. The optimization with respect to $\sigma_{L^2}$ means adding and removing mass locally at a given point in the domain. On the other hand, the optimization with respect to $\sigma_{PC}$ is driven by the shape derivative flux and moves only the interface $\partial D$.

The regularization functional (5.8) becomes

$$\mathcal{R}_{U,W}(\sigma_{PC}, \sigma_{L_2}, \lambda) = (1-\lambda)\beta \left\|\sigma_{L^2}\right\|^2_{L^2(\Omega)} + \lambda\alpha \left\|\sigma_{PC}\right\|^2_{BV(\Omega)}. \tag{6.8}$$

The $\mathcal{R}_W = \|\cdot\|^2_{L^2(\Omega)}$ trivially fulfills the assumption (A4). The assumption (A2) is dependent on the specific forward problem. For magnetic induction tomography it will

[2] In our case, the "shape" of the piecewise constant $\sigma$ defined by (6.1), the topology is determined by the number of connected components of $D$ and their equivalent classes (ball, torus etc.).

be established in Section 6.2. The problematic assumptions are (A1) and (A3). First, the space $BV(\Omega)$ is not reflexive. A direct remedy is to approximate $BV(\Omega)$ by its reflexive subspace $W^{1+\eta}(\Omega)$, $0 < \eta \ll 1$, which resolves also the non-differentiability of the $BV-$norm. The second possibility is to follow the analysis in [1]. There, the convergence in $BV(\Omega)$ is understood in the weaker then norm topology sense, namely in $L^p-$sense [3]. The seminorm $J(\sigma)$ in $BV(\Omega)$ is furthermore efficiently approximated by the functional ([1, Theorem 2.2])

$$J_\varepsilon(\sigma) = \sqrt{|\nabla\sigma|^2 + \varepsilon}, \qquad \varepsilon > 0, \tag{6.9}$$

which is differentiable everywhere. We note that $\varepsilon$ will be used subsequently in different situations and it always represents a small positive number.

We conclude that for the admissible forward operator $F$ the topology-to-shape continuation method lies within the proposed continuation approach (Chapter 5).

Despite all the effort in combining topology and shape sensitivity concepts and some very positive results as stated in Section 6.1.1, no clear idea was yet presented how these concepts could be unified in one approach. We quote [50]:"It is still an open problem to devise how the combination of boundary variations and singular perturbations of geometrical domains enters in a general approach of shape optimization." We think the idea of continuation extension of Tikhonov regularization presented in this article is such a approach. But, we view the problem from a different angle. We first identify the optimal distribution of the unknown parameter which represents the topology. We then continuously recast this information to the optimal shape without using any singular perturbations of the geometry. The difficulties in combining the local and global sensitivity concepts vanish. Even though the numerical experiments show promising results, we have no proof of the global convergence. We remark, that the approach of singular perturbations of the geometry e.g. in [3] allow to adapt the topology during the algorithm. It is thus more general.

Let us quote also from [111], where an penalty method is used to solve piecewise constant parameter identification problems: "From our numerical experiences, we find that it is better to neglect the regularization term at the beginning stage of the iteration. At this stage, we should let the output-least-squares term to drag $\phi$[4] into the right direction without thinking about the regularity of $q$[5]." In the context of continuation it is easy to explain this observation from [111]. The minimization without total variation regularization term essentially behaves as Landweber type of regularization method, where the number of iterations plays the role of regularization [42], and the method converges to the least square solution in $L^2-$sense. Gradually increasing regularization parameter in the front of the total variation term functions as the continuation parameter $\lambda$. The same insight explains the global convergence of augmented Lagrangian methods [37].

[3] Interestingly, it is the topology of $W$.
[4] piecewise constant level set function
[5] coefficient to be recovered

The advantage of the continuation approach is that the relaxed space $W$ does not have to be $L^2(\Omega)$.

## 6.2 Magnetic induction tomography

In this section we apply the approach to an inverse problem in magnetic induction tomography (MIT). The inverse problem consists in recovering a piecewise-constant parameter from boundary data.

Magnetic induction tomography is a non-invasive visualization technique. It is a very promising member of the broader electromagnetic imaging family with biomedical and industrial applications, for instance non-destructive testing, industrial and medical imaging [55]. We refer the reader to the paper [108] for a comprehensive review of mathematical methods in electromagnetic tomography techniques. Magnetic induction tomography is a non-contact technique, in contrast to widely studied electrical impedance tomography [29, 17]. The errors caused by the electrode/body contact can be avoided completely. Another advantage of MIT is its explicit frequency dependence which allows for more accurate reconstruction of the body properties [21].

### 6.2.1 Mathematical formulation

We proceed to the mathematical description of MIT. Electromagnetic phenomena in general are governed by the Maxwell equations. For the linear isotropic case in the time-harmonic regime with the angular velocity $\omega > 0$, they take form

$$\begin{aligned} \nabla \times \mu^{-1}\boldsymbol{B} &= i\omega\epsilon\boldsymbol{E} + \boldsymbol{J}, \\ \nabla \times \boldsymbol{E} &= -i\omega\boldsymbol{B}, \\ \nabla \cdot \boldsymbol{B} &= 0, \quad \nabla \cdot \epsilon\boldsymbol{E} = 0, \end{aligned} \tag{6.10}$$

where $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic field, respectively. The permeability $\mu$ and the permitivity $\epsilon$ are known strictly positive scalar functions of the space variable. By Ohm's law, the divergence-free current $\boldsymbol{J}$ is the sum of the applied current $\boldsymbol{J}_e$ from the excitation coil and the induced current $\sigma\boldsymbol{E}$. The conductivity $\sigma$ is assumed to be positive in the imaged body and it vanishes in the surrounding non-conducting region. Making use of the magnetic vector potential $\boldsymbol{A} = \nabla \times \boldsymbol{B}$, we can reformulate the system (6.10) as

$$\begin{aligned} \nabla \times (\mu^{-1}\nabla \times \boldsymbol{A}) + i\omega(\sigma + i\omega\epsilon)\boldsymbol{A} &= \boldsymbol{J}_e, \\ \nabla \cdot (\epsilon\boldsymbol{A}) &= 0, \end{aligned} \tag{6.11}$$

if the scalar potential $V$ is fixed by the temporal gauge, i.e. $V = 0$. The equations above have to be accompanied by appropriate boundary conditions. For more on various MIT models we refer to [129, 108].

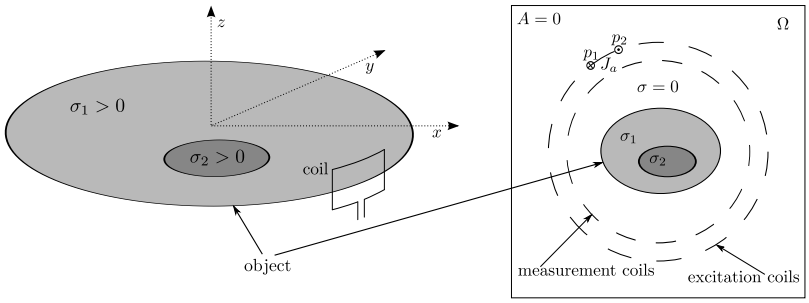



Figure 6.1: Magnetic induction tomography setup

We formulate a simplified MIT boundary value problem. Let $\Omega$ be a bounded two-dimensional domain in the $xy$-plane with the sufficiently smooth boundary $\partial\Omega =: \Gamma$. It represents a cross section of the imaged body. Assume that the applied current $\boldsymbol{J}_e$ is perpendicular to $xy$-plane and does not depend on $z$-coordinate. The induced eddy currents can be then described by the $z$-component of the potential $\boldsymbol{A}$, which we will simply denote by $A$. Accurate modeling of the excitation coils generating $\boldsymbol{J}_e$ is a very challenging task on its own [56]. We restrict ourselves to the imaged body region, where the conductivity is strictly positive, $\sigma \geq \sigma_{\min} > 0$. The domain source $\boldsymbol{J}_e$ is modeled by an boundary source $e$, which is imposed via the Neumann boundary condition on $\Gamma$. The corresponding experimental setup is depicted in Figure 6.1. Contact-free excitation coils induce a magnetic field, which penetrates the imaged body. The electromagnetic response is then measured by detection coils as voltages. For an experimental realization see [75].

We use the eddy current approximation of the Maxwell equations, where the displacement current term $i\omega\epsilon\boldsymbol{A}$ in (6.11) is disregarded. The state variable $A$ then satisfies the forward problem

$$\begin{aligned} \nabla \cdot (\mu^{-1}\nabla A) + i\omega\sigma A &= 0 \qquad \text{in } \Omega, \\ \mu^{-1}\nabla A \cdot \boldsymbol{n} &= e \qquad \text{on } \Gamma. \end{aligned}$$

Let us remark, that under physiological conditions for higher excitation frequencies $\omega$ the displacement current term can have a significant contribution and has to be taken into account.

### 6.2.2 Forward problem

We now show that the MIT forward problem satisfies the assumption (A2) of Section 5.2.

Let us first introduce some notation. The standard scalar product of two complex valued functions in the space $L^2(\Omega)$ is denoted by $(u, v) = \int_\Omega u(x)\overline{v(x)}\,\mathrm{d}x$. We write $\|u\| = \sqrt{(u, u)}$ for the induced norm. The subscript $\Gamma$ indicates integration over the boundary in $L^2(\Gamma)$-sense. The symbol $H^1(\Omega)$ stands for the Sobolev space of the complex-valued functions with first weak derivatives. It is compactly embedded in the all Lebesgue spaces but $L^\infty(\Omega)$ (e.g. [77, Theorem 5.8.2]):

$$H^1(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega) \quad \text{for any } q \in [1, \infty). \tag{6.12}$$

The weak formulation of (6.12) reads as

$$\left(\mu^{-1}\nabla A, \nabla\varphi\right) + (i\omega\sigma A, \varphi) = (e, \varphi)_\Gamma \qquad \forall\varphi \in H^1(\Omega). \tag{6.13}$$

This variational problem defines the *impedance map* $\Lambda$, the so-called Neuman-to-Dirichlet map

$$\Lambda : (\sigma, \omega, e) \mapsto A|_\Gamma. \tag{6.14}$$

**Lemma 6.1.** *The impedance map*

$$\Lambda : \sigma \mapsto \Lambda(\sigma) = A|_\Gamma,$$

*where the function $A$ is the solution of the problem (6.13) for any $e \in L^2(\Gamma)$ and $\omega > 0$ fixed, is a well-defined and strongly continuous map from the set*

$$M = \{\sigma \in L^q(\Omega), q > 1 : \sigma \geq \sigma_{\min} > 0\}.$$

*to the space $L^2(\Gamma)$.*

*Proof.* The Sobolev embedding (6.12) implies that the term in (6.13) containing $\sigma$ makes sense for any $\sigma \in L^q(\Omega)$, $q > 1$. Given arbitrary $\sigma \in M$, the existence of a unique solution $A \in H^1(\Omega)$ follows readily from the Lax-Milgram theorem for sesquilinear forms (Theorem A.5).

Following the standard techniques, we set $\varphi = A$ in (6.13) to obtain

$$\frac{\min\{{\mu^{-1}}_{\min}, \omega\sigma_{\min}\}}{C} \|A\|^2_{H^1(\Omega)} \leq \|e\|_\Gamma \|A\|_\Gamma,$$

and so we have the a priori estimate

$$\|\Lambda(\sigma)\|_\Gamma \leq \|A\|_{H^1(\Omega)} \leq C \|e\|_\Gamma.$$

Let now $\sigma_n \rightharpoonup \sigma$ as $n \to \infty$. It holds that $\sigma \in M$, because $M$ is closed and convex. Denote by $A_n$ and $A$ the corresponding solutions of (6.13) for $\sigma_n$ and the weak limit $\sigma$, respectively. The subtraction of the variational formulas from each other gives

$$\left(\mu^{-1}\nabla(A - A_n), \nabla\varphi\right) + (i\omega\sigma(A - A_n), \varphi) = (i\omega(\sigma_n - \sigma)A, \varphi).$$

The sesquilinear form on the left hand side is equivalent to the $H^1(\Omega)$-scalar product which leads to a one-to-one correspondence between test functions $\varphi$ and linear functionals on $H^1(\Omega)$. Since $A\varphi \in L^{q/(q-1)}(\Omega)$, the right hand side tends to zero for any $\varphi \in H^1(\Omega)$ as $n \to \infty$. Hence, we see that

$$A_n \rightharpoonup A \quad \text{in } H^1(\Omega).$$

It follows from the continuity of the trace mapping $H^1(\Omega) \to H^{1/2}(\Gamma)$ and the compact embedding $H^{1/2}(\Gamma) \hookrightarrow\hookrightarrow L^2(\Gamma)$, that

$$A_n \to A \quad \text{in } L^2(\Gamma).$$

□

The differentiation of (6.13) at $\sigma$ in the direction $h$ yields

$$(\mu^{-1}\nabla\delta A, \nabla\varphi) + (i\omega\sigma\delta A, \varphi) = -(i\omega h A, \varphi) \qquad \forall\varphi \in H^1(\Omega). \tag{6.15}$$

The symbol $\delta A := \delta A(\sigma; h)$ stands for the *variation* (Gâteaux differential) of $A = A(\sigma)$ in the direction $h$. The variation $\delta A$ is sometimes called the sensitivity of $A$ and (6.15) the *sensitivity equation*, which is a well-posed problem with the unique solution $\delta A$ for any $h$ from $L^q(\Omega), q > 1$ satisfying the a priori estimate

$$\|\delta A\|_{H^1(\Omega)} \le \frac{C}{\min\{{\mu^{-1}}_{\min}, \sigma_{\min}\}} \|h\|_{L^q(\Omega)} \|A\|_{L^{q/(2(q-1))}(\Omega)}.$$

It is straightforward to verify that for given $\sigma$ the mapping $h \mapsto \delta A(\sigma; h)|_\Gamma$ is a linear and bounded operator with

$$\begin{aligned}\|\delta A(\sigma;\cdot)\|_{L(M,L^2(\Gamma))} &= \sup_{\|h\|_{L^q(\Omega)}=1} \frac{\|\delta A(\sigma; h)\|_{L^2(\Gamma)}}{\|h\|_{L^q(\Omega)}} \\ &\le \frac{C}{\min\{{\mu^{-1}}_{\min}, \sigma_{\min}\}} \|A\|_{L^{q/(2(q-1))}(\Omega)}.\end{aligned}$$

Recalling the relationship between the variation and Fréchet derivative (Theorem A.6), we see that $\Lambda$ is Fréchet differentiable and

$$\Lambda'(\sigma)h = \delta A(\sigma; h)|_\Gamma.$$

The map $\Lambda' : M \to L(M, L^2(\Gamma))$ is continuous in $\sigma$ by the similar reasoning as in the proof of Theorem 6.1 and so we have the following assertion.

**Lemma 6.2.** *The impedance map $\Lambda : M \to L^2(\Gamma)$ is $C^1$-Fréchet differentiable.*

### 6.2.3 Inverse problem

By the inverse problem in MIT, we will understand the reconstruction of the conductivity $\sigma$ in the imaged body based on a finite number of Dirichlet-to-Neumann data $(e, m)$ corresponding to the impedance map (6.14). The boundary data $m$ are essentially voltage measurements associated with the excitations $e$. Lemma 6.1 implies that $\Lambda$ is a completely continuous operator and so the recovery of $\sigma$ is inherently an ill-posed problem.

We employ the topology-to-shape continuation method (TSCM) from Section 6.1.2 to solve MIT. We look for the conductivity $\sigma$ in the form (6.7), i.e.

$$\sigma = (1-\lambda)\sigma_{L^2} + \lambda\sigma_{PC},$$

where $\sigma_{PC}$ is a double-valued piecewise constant function as it is considered in Section 6.1 for the example of electrical impedance tomography. The associated continuation Tikhonov functional for MIT read as

$$\mathcal{T}_{\alpha,\beta}(\sigma) = \mathcal{F}(\sigma) + \mathcal{R}_{U,W}(\sigma_{PC}, \sigma_{L_2}, \lambda), \tag{6.16}$$

where $\mathcal{F}$ is the fidelity term

$$\mathcal{F}(\sigma) = \int_\Gamma |\Lambda(\sigma, \omega, e) - m|^2 \,\mathrm{d}S. \tag{6.17}$$

The regularization part $\mathcal{R}_{U,W}$ is given by

$$\mathcal{R}_{U,W}(\sigma_{PC}, \sigma_{L_2}, \lambda) = \lambda\alpha\left[\|\sigma_{PC}\|^2 + \sqrt{|\nabla\sigma_{PC}|^2 + \varepsilon}\right] + (1-\lambda)\beta\|\sigma_{L^2}\|^2, \tag{6.18}$$

which complies with the TSCM analysis in Section 6.1.2. The forward problem operator $\Lambda$ of MIT is an admissible operator fulfilling assumption (A2) of Section 5.2 as it is shown in Section 6.2.2. Altogether, the theory of Section 5.2 is applicable to the inverse problem of MIT as stated in this section.

#### Adjoint problem

In Section 6.2.5 we will use a gradient-based method (the steepest descent method) to find a minimizer of (6.16). Let us express the derivative of fidelity term (6.17) using an *adjoint variable*. The variation of $\mathcal{F}$ in the direction $h$ reads as

$$\begin{aligned}\delta\mathcal{F}(\sigma; h) &= \lim_{t\to 0}\frac{\mathcal{F}(\sigma+th) - \mathcal{F}(\sigma)}{t} \\ &= (\Lambda(\sigma) - m, \delta\Lambda(\sigma; h))_\Gamma + (\delta\Lambda(\sigma; h), \Lambda(\sigma) - m)_\Gamma \\ &= 2\Re\left[(\delta\Lambda(\sigma; h), \Lambda(\sigma) - m)_\Gamma\right],\end{aligned}$$

where the variation $\delta\Lambda(\sigma;h) \equiv \delta A$ solves the sensitivity equation (6.15) and $\Re$ returns the real part of a complex number. We now introduce the adjoint variable $Z$, which satisfies

$$(\mu^{-1}\nabla\varphi, \nabla Z) + (i\omega\sigma\varphi, Z) = -(\varphi, \Lambda(\sigma) - m)_\Gamma \qquad \forall\varphi \text{ in } H^1(\Omega), \tag{6.19}$$

to establish that

$$\begin{aligned} \delta\mathcal{F}(\sigma;h) &= 2\Re\left[(\delta\Lambda(\sigma;h), \Lambda(\sigma) - m)_\Gamma\right] \\ &\overset{(6.19)}{=} 2\Re\left[-(\mu^{-1}\nabla\delta A, \nabla Z) - (i\omega\sigma\delta A, Z)\right] \\ &\overset{(6.15)}{=} 2\Re\left[(i\omega h A, Z)\right]. \end{aligned} \tag{6.20}$$

Let us note, that the variational problem (6.19) for $Z$ is uniquely solvable given the properties of the material parameters and of the impedance map $\Lambda$. We assume that $m \in L^2(\Gamma)$.

### 6.2.4 Implementation of TSCM method

In this section we describe the implementation of topology-to-shape continuation method (TSCM) for the problem of the magnetic induction tomography.

The practical implementation of the TSCM algorithm presented in Algorithm 1 closely follows the theoretical exposition. The outside cycle successively increases value of $\lambda$ by $\Delta\lambda$ starting from $\lambda = 0$. It terminates when $\lambda = 1$ is reached. The number of steps is determined by $\Delta\lambda$. The inner cycle constitute more or less a standard adjoint-variable based steepest descent algorithm for minimization of (6.16) for the fixed $\lambda$. The number $n$ stands for the total number of iterations through both loops in Algorithm 1.

We use the level set method to parametrize the piecewise-constant conductivity $\sigma_{PC}$, see (6.1). This method was originally proposed in [89] for analyzing the motion of an interface. Since then it has found many applications in very diverse fields. The main idea lies in defining the level set function $\phi$ for the subset $D \subset \Omega$ with its boundary $\partial D$

$$\phi(x) = \begin{cases} \text{distance}(x, \partial D) & x \in D, \\ -\text{distance}(x, \partial D) & x \in \Omega/D. \end{cases}$$

The zero level set of $\phi$ represents the boundary of $D$ (its "interface"). The piecewise-constant conductivity $\sigma_{PC}$ is then parametrized as

$$\sigma_{PC}(\phi) = \sigma_1 H(\phi) + \sigma_2(1 - H(\phi)). \tag{6.21}$$

The symbol $H$ stands for Heaviside step function

$$H(\phi) = \begin{cases} 1 & \text{if } \phi > 0, \\ 0 & \text{if } \phi < 0. \end{cases}$$

**Data**: $\lambda = 0;\ \sigma = \sigma_{L^2} = \delta_1;\ \phi = -\delta_2;\ n = 0$
**do**
  $s_n = 2$;
  **do**
    *Compute the derivatives:*
      $\sigma_n \longrightarrow$ direct problem (6.13) $\longrightarrow A_n$;
      $(\sigma_n, A_n) \longrightarrow$ adjoint problem (6.19) $\longrightarrow Z_n$;
      $(A_n, Z_n) \longrightarrow$ cost functional derivative (6.20) $\longrightarrow \nabla_\sigma \mathcal{F}_n$;
      $\nabla_\sigma \mathcal{F}_n + (6.26) + (6.25) \longrightarrow \nabla_{\sigma_{L^2}} \mathcal{T}_{\alpha,\beta,n}$;
      $\nabla_\sigma \mathcal{F}_n + (6.27) + (6.24) \longrightarrow \nabla_\phi \mathcal{T}_{\alpha,\beta,n}$;

    *Find the optimal step:*
      $s_n = \text{Linesearch}(\sigma_n, \nabla_{\sigma_{L^2}} \mathcal{T}_{\alpha,\beta,n}, \nabla_\phi \mathcal{T}_{\alpha,\beta,n})$;
    *Update $\sigma_n$:*
      $\sigma_{L^2,n+1} = \sigma_{L^2,n} - s_n \nabla_{\sigma_{L^2}} \mathcal{T}_{\alpha,\beta,n}$;
      $\phi_{n+1} = \phi_n - s_n \nabla_\phi \mathcal{T}_{\alpha,\beta,n}$;
      $\sigma_{n+1} = \lambda \sigma_{PC}(\phi_{n+1}) + (1-\lambda)\sigma_{L^2,n+1}$;

    $n = n + 1$;
  **while** $|\nabla_{\sigma_{L^2}} \mathcal{T}_{\alpha,\beta,n}|^2 + |\nabla_\phi \mathcal{T}_{\alpha,\beta,n}|^2 > \tau_1^2$ and $s_n > \tau_2$;
  $\lambda = \lambda + \Delta\lambda$;
**while** $\lambda < 1$;

Algorithm 1: Topology-to-shape continuation algorithm

We use the following smooth approximation $H_\varepsilon$ and its derivative

$$H_\varepsilon(\phi) = \frac{1}{\pi} \arctan \frac{\phi}{\varepsilon} + \frac{1}{2}, \qquad H'_\varepsilon(\phi) = \delta_\varepsilon(\phi) = \frac{\varepsilon}{\pi(\phi^2 + \varepsilon^2)} \tag{6.22}$$

to avoid the difficulties connected with the non-differentiability of the original function $H$.

The gradient of (6.18) with respect to $\sigma_{PC}$ is evaluated as the solution of the variational problem

$$(\nabla_{\sigma_{PC}} \mathcal{R}_{U,W}, h) = \lambda\alpha \left[ \left( \frac{\nabla \sigma_{PC}}{\sqrt{|\nabla \sigma_{PC}|^2 + \varepsilon}}, \nabla h \right) + (\sigma_{PC}, h) \right] \tag{6.23}$$

for all $h \in H_0^1(\Omega)$. All the variational problems ((6.13), (6.19) etc.) are solved by the finite element method where $H^1(\Omega)$ is approximated by linear Lagrange basis functions. Solving (6.23) means we project $\partial_{\sigma_{PC}} \mathcal{R}_{U,W}$ onto the nodes of the finite element mesh.

Using (6.21) together with (6.22) we have

$$\nabla_\phi \mathcal{R}_{U,W} = (\sigma_1 - \sigma_2) H'_\varepsilon(\phi) \nabla_{\sigma_{PC}} \mathcal{R}_{U,W}. \tag{6.24}$$

The gradient of (6.18) with respect to $\sigma_{L^2}$ is simply

$$\nabla_{\sigma_{L^2}} \mathcal{R}_{U,W} = 2(1-\lambda)\beta\sigma_{L^2}. \tag{6.25}$$

The gradient $\nabla_\sigma \mathcal{F}$ of the fidelity term $\mathcal{F}$ with respect to $\sigma$ is evaluated from (6.20) again by projection onto the nodes of the finite element mesh as in (6.23):

$$\nabla_\sigma \mathcal{F} = 2\Re[i\omega A Z].$$

This yields

$$\nabla_{\sigma_{L^2}} \mathcal{F} = (1-\lambda)\nabla_\sigma \mathcal{F} \tag{6.26}$$

and

$$\nabla_\phi \mathcal{F} = \lambda(\sigma_1 - \sigma_2) H'_\varepsilon(\phi) \nabla_\sigma \mathcal{F}. \tag{6.27}$$

We do not optimize with respect to the constants $\sigma_1$ and $\sigma_2$, which we consider to be known. However, Algorithm 1 is easily extendable to the case of unknown $\sigma_1$ and $\sigma_2$.

We emphasize that we do not assume any a priori knowledge about the shape of $D$. The unknowns $\phi$ and $\sigma_{L^2}$ are initiated as $\phi = -\delta_1$ and $\sigma_{L^2} = \delta_2$ with $\delta_1$ and $\delta_2$ being some positive constants, $\delta_2 \approx \sigma_{\min}$. It means that initially ($\lambda = 0$) the whole domain $\Omega$ is occupied by a weak phase. In addition we have zero inclusion $D$ and thus the value of $\sigma_{PC} \equiv \sigma_2$ in the whole domain.

In Algorithm 1 the search for an optimal step-size $s_n$ might be the most time-consuming part, since the Linesearch-algorithm detects the optimal $s_n$ by the evaluation of the cost functional for different intermediate values of $s_n$ and one such evaluation means to solve one forward problem (6.13). However, we do not need to find the optimal value of $s_n$ for which the drop of $\mathcal{T}_{\alpha,\beta}$ is maximal. It is enough to find one value for which $\mathcal{T}_{\alpha,\beta}$ drops sufficiently (the method is then no more steepest descent). We update $s$ according to the following simple rule [33]:

$$s_{n+1} = 2s_n \text{ if } \mathcal{T}_{\alpha,\beta}(\sigma_n(s_{n-2})) < \mathcal{T}_{\alpha,\beta}(\sigma_{n-1}),$$

i.e. when $s_{n-1} := s_{n-2}$ gave a reduction of cost functional value, we try double the step. If in the next step $s_n$ does not give a descent, we take the step with the smallest $k$ from the sequence $s_n^k = s_n^{k-1}/2$, $k = 1, \ldots, \infty$ such that we have descent. The last part is the actual update process. The inner cycle of Algorithm 1 stops when the norm of gradient is sufficiently small ($\leq \tau_1$) or the computed gradient is not a descent direction anymore, i.e. $s_n \leq \tau_2$, where $\tau_1$ and $\tau_2$ are suitable constants.

### 6.2.5 Numerical experiments

In all the experiments we use synthetic data. The number $N$ of the measurements for every experiment corresponds to the number of excitation coils $N(e)$ (see Figure 6.1) multiplied with the number of excitation frequencies $N(\omega)$. The fidelity functional reads

$$\mathcal{F}(\sigma) = \sum_{\omega} \sum_{e} \int_{\Gamma} |\Lambda(\sigma, \omega, e) - m_{\omega,e}|^2 \, \mathrm{d}S. \tag{6.28}$$

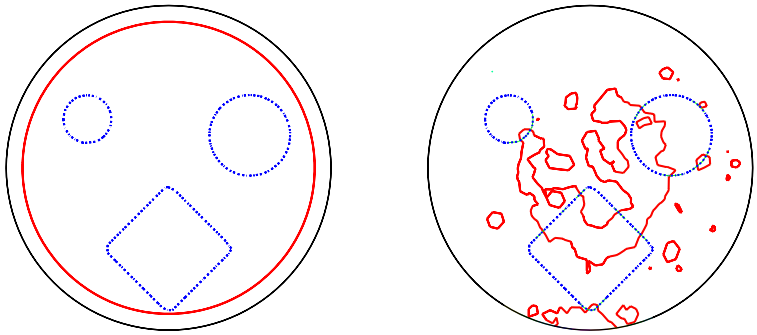
We take $\sigma_1 = 20S \cdot m^{-1}$ and $\sigma_2 = 2S \cdot m^{-1}$ and $\mu = \mu_0$ which complies with physiological conditions. For comparison, in non-destructive testing of metallic pieces normal magnitudes of $\sigma$ are in millions of $S \cdot m^{-1}$ and $\mu \gg \mu_0$.

All the excitation currents $e_i = 1A \cdot m^{-1}$, $i = 1, \ldots, N(e)$. The angular excitation frequencies $\omega_i = 2\pi f_i = 2\pi 2^{15+i}$, $i = 0, \ldots, N(\omega) - 1$. The basic frequency $f_0 = 2^{15}$ is set so that $\mu^{-1} > \omega_0 \max(\sigma_1, \sigma_2)$. For such a base frequency the magnetic phenomena dominate the electric ones.

The parameters in Algorithm 1 are $\tau_1 = 10^{-5}$, $\tau_1 = 10^{-6}$, $\delta_1 = 1$, $\delta_2 = 0.01$. We implemented the algorithm in FreeFem++ [61]. In all the experiments for both $\sigma_{L^2}$ and $\phi$ we use identical fixed regular meshes with homogeneous division of the boundary $\Gamma$. We also always consider 28 excitation coils, i.e. $N(e) = 28$, and the regularization parameters $\alpha$ and $\beta$ are fixed as $\alpha = \beta = 0.00001$. In (6.22) we take $\epsilon = h^2$, where $h$ is the diameter of the finite element mesh. We take $\Delta\lambda = 0.1$.

We first compare the performance of the continuation algorithm (TSCM) and the standard level set method (LSM) on an example with a non-trivial topology (Figure 6.2). The blue dotted line represents in all the figures the exact phantom and the red line is the numerical approximation. The initial shape of $\sigma_{PC}$ for the standard LSM is depicted in Figure 6.2(a). Figure 6.2 displays the results for the base angular frequency $\omega_0$. The LSM in Figure 6.2(a) ended up in a local minimum after 37 iterations. The algorithm stopped because the computed gradient was not a descent direction anymore, i.e. $s_{37} < \tau_2$. We see that without a proper initial guess, the standard LSM failed to recover the desired shape. On the other hand, the TSCM in Figure 6.2(c) for zero noise provided a decent approximation. Both bigger phantoms are recovered quite successfully but they stay connected. The smallest phantom is not identified properly. Only certain allocation of its mass is identified along the proximal boundary. Even for $1\%$ noise the TSCM method provided a decent approximation (Figure 6.2(d)). The method seems to be rather stable with respect to noise. We recall, that the standard LSM is very sensitive when only boundary measurements are available, e.g. in [30, Figure 7] only a noise level of $0.01\%$ is considered in a case of a complicated phantom for the problem of electric impedance tomography.

We next perform numerical experiments that use explicit dependency of MIT model on the frequency $\omega$. The results are presented in Figure 6.3 for the phantom identical to the previous single-frequency experiment in Figure 6.2. We consider the four-frequency



(a) LSM, no noise, initial $\phi_0$

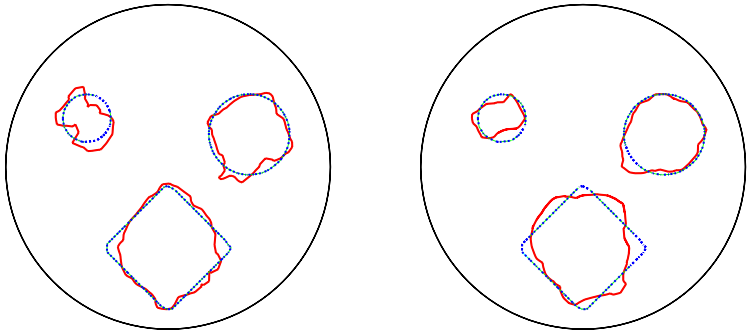
(b) LSM, no noise, final $\phi_{37}$

(c) TSCM, no noise, final $\phi_{228}$

(d) TSCM, $1\%$ noise, final $\phi_{188}$

Figure 6.2: Comparision between the standard LSM and TSCM

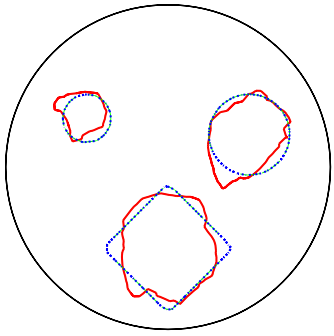
case $N(\omega) = 4$ and four levels of noise: $1\%$, $5\%$, $10\%$ and $20\%$. The blue line is again the exact shape and the red line is its TSCM-identification. As expected we got more accurate recovery of $\sigma_{PC}$. For the noise levels up to $10\%$ all the components of the phantom are quite accurately identified, accuracy gradually decreasing. Even for noise level of $20\%$, the identification is surprisingly accurate and all the components are identified, however two bigger components stay connected by a bridge. This experiment confirms our conjecture that the method is very stable with respect to the non-systematic noise.



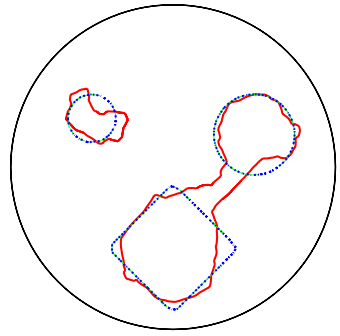
(a) $1\%$ noise, final $\phi_{454}$



(b) $5\%$ noise, final $\phi_{350}$



(c) $10\%$ noise, final $\phi_{290}$



(d) $20\%$ noise, final $\phi_{386}$

Figure 6.3: TSCM: multiple frequency case $N(\omega) = 4$

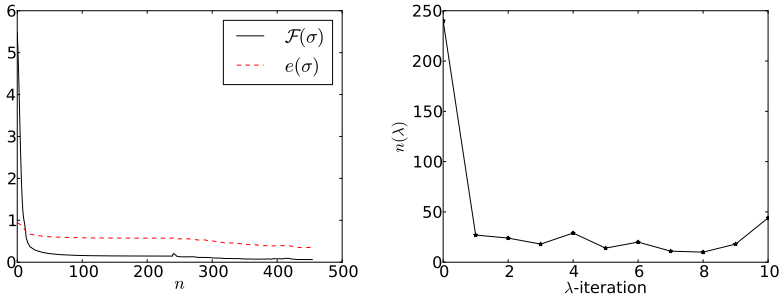
Noise causes non-convexity of the fidelity term $\mathcal{F}$ regardless the properties of the forward operator $F$. Provided the data contain sufficient information to identify the phantom, the TSCM is able to eliminate this type of non-convexity. We are convinced the reason lies within the nature of the method. The TSCM is essentially a convexification

approach.

The convergences of the fidelity term $\mathcal{F}(\sigma)$ and of the relative error between the computed conductivity $\sigma_{TSCM}$ and exact conductivity $\sigma_{exact}$

$$e(\sigma) = \frac{\|\sigma_{TSCM} - \sigma_{exact}\|_{L^2(\Omega)}}{\|\sigma_{exact}\|_{L^2(\Omega)}} \tag{6.29}$$

with respect to the total number of iterations $n$ of Algorithm 1 are depicted in Figure 6.4(a). These graphs correspond to the experiment of Figure 6.3(a). The distribution of the number of iterations for different $\lambda-$steps is depicted in Figure 6.4(b). In general, the first iteration of the TSCM for $\lambda = 0$ is the most time consuming, which is natural, because it is nothing else than the minimization of $\mathcal{T}_{\alpha,\beta}$ in the space $L^2(\Omega)$. It provides the information about "the optimal topology" for $\sigma_{PC}$. Once this good initial guess is found, the continuation method rather quickly transforms this function to the desired piecewise-constant conductivity $\sigma_{PC}$.




(a) Convergence (b) Distribution of iterations

Figure 6.4: TSCM: experiment 1; $\rho = 1\%$; $N(\omega) = 4$

# Conclusions

In this thesis, we have discussed a few numerical techniques in partial differential equations.

In the first part of the thesis, we have considered nonlinear degenerate parabolic boundary value problems with nonstandard boundary conditions. We have proved the well-posedness of these problems and convergence of the time and the space discretization schemes. Our analysis relies on Rothe's method and monotone operator theory.

The primary motivation for studying boundary value problems with the nonstandard boundary conditions comes from the mathematical modelling of various physical phenomena. Unlike the standard boundary conditions, the nonstandard ones can take into account sometimes very complicated dynamics of the various processes on the boundary. This has been illustrated in Chapter 2, where a dynamical boundary condition has been used to model the rainfall infiltration through soil. In Chapter 3 we have generalized the theoretical results to a broader class of nonstandard boundary conditions. Chapter 4 deals with nonlinear eddy current problem. A possible extension here is to incorporate a surface rotor operator into the impedance-like boundary condition, which would be analoguous to the boundary condition from Chapter 3. In this way, one could model surface currents on the boundary.

The second part of the thesis presents the results of the collaboration with Valdemar Melicher. We have proposed a continuation approach for Tikhonov regularization and employed it to perform shape identification without any initial knowledge of topology. We have successfully applied the resulting topology-to-shape continuation method (TSCM) to a magnetic induction tomography (MIT) problem.

This method appears to be a very promising candidate for an ultimate framework unifying both topology and shape sensitivities. To establish such a claim more rigorously, it is necessary to provide a deeper analysis of the continuation approach with respect to the homotopy parameter $\lambda$, which is a possible future work. Any result in this direction will be dependent on a particular choice of the functional spaces $W$ and $U$ and their properties. Our understanding of the underlying concepts suggests that for TSCM-specific choice of the functional spaces such an analysis is attainable.

Possible future work with respect to the TSCM or to the continuation approach in

general is to propose and analyze appropriate parameter choice rules (PCRs) for the two regularization parameters $\alpha$ and $\beta$ in (5.6). The regularization parameters could be considered as functions of $\lambda$ as well. This should lead to $\lambda$-adaptive PCRs and consequently a more efficient implementation of the TSCM algorithm. From the numerical point of view also conjugate gradient, quasi-Newton or Gauss-Newton algorithm extensions are possible.

# Part III

# Appendices

# Appendix A

# Appendix

For the convenience of the reader, this appendix contains important formulas, definitions and assertions. We first state a few practical mathematical inequalities. We then recall basic notions of functional analysis and variational calculus. We formulate the main theorem on monotone operators and provide a sketch of the proof. Afterwards we briefly review Sobolev spaces and the appendix ends with a short description of finite element method.

## A.1 Useful (in)equalities

**Young's inequality.** Suppose that $a$ and $b$ are nonnegative real numbers and $p$ and $q$ are positive real numbers such that $1/p + 1/q = 1, p \geq 1$. Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

**Young's inequality with $\varepsilon$.** Suppose that $a$ and $b$ are nonnegative real numbers and $p$ and $q$ are positive real numbers such that $1/p + 1/q = 1, , p \geq 1$. Then

$$ab \leq \varepsilon a^p + C_\varepsilon b^q,$$

where $\varepsilon$ is arbitrarily small and $C_\varepsilon$ arbitrarily large.
**Hölder's inequality - discrete version.** Suppose $n \in \mathbb{N}$, $p > 1$ and $1/p + 1/q = 1$. If $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are nonnegative real numbers, then

$$\sum_{i=1}^{n} a_i b_i \leq \left( \sum_{i=1}^{n} a_i^p \right)^{1/p} \left( \sum_{i=1}^{n} b_i^q \right)^{1/q}.$$

**Hölder's inequality - integral version.** Suppose $d \in \mathbb{N}$, $p > 1$ and $1/p + 1/q = 1$. If $f$ and $g$ are integrable functions on $\Omega \subset \mathbb{R}^d (\mathbb{C}^d)$, then

$$\int_\Omega |f(x)g(x)| \, \mathrm{d}x \leq \left( \int_\Omega |f(x)|^p \, \mathrm{d}x \right)^{1/p} \left( \int_\Omega |g(x)|^q \, \mathrm{d}x \right)^{1/q}.$$

The special case $p = q = 2$ gives a form of the Cauchy-Schwarz inequality.

**Jensen's inequality - discrete version.** Let $\phi$ be a convex function on the real line and $a_i, \ldots, a_n, n \in \mathbb{N}$ real numbers. Then

$$\phi \left( \frac{\sum_{i=1}^n a_i}{n} \right) \leq \frac{\sum_{i=1}^n \phi(a_i)}{n}.$$

**Jensen's inequality - integral version.** Let $\phi$ be a convex function on the real line and $f$ a real valued integrable function on $\Omega$. Then

$$\phi \left( \int_\Omega f(x) \, \mathrm{d}x \right) \leq \int_\Omega \phi(f(x)) \, \mathrm{d}x.$$

**Gronwall's inequality - discrete version.** Let $\{a_n\}$ and $\{A_n\}$ be sequences of nonnegative real numbers and $q > 0$. If

$$a_n \leq A_n + \sum_{i=1}^{n-1} q a_i \qquad \forall n \in \mathbb{N},$$

then

$$a_n \leq A_n + e^{nq} \sum_{i=1}^{n-1} q A_i \qquad \forall n \in \mathbb{N}.$$

**Gronwall's inequality - continuous version [11]** Let $y$ and $h$ be continuous real functions on the interval $[a, b] \subset \mathbb{R}$. Suppose that $h$ is nonnegative and nondecreasing, and $C$ is a positive constant. If

$$y(t) \leq h(t) + C \int_a^t y(s) \, \mathrm{d}s \qquad \forall t \in [a, b],$$

then

$$y(t) \leq h(t) e^{C(t-a)} \qquad \forall t \in [a, b].$$

**Abel's summation rule.** For every set of real numbers $a_0, \ldots, a_n$ it holds true that

$$\sum_{i=1}^n (a_i - a_{i-1}) a_i = \frac{1}{2} \left( a_n^2 - a_0^2 + \sum_{i=1}^n (a_i - a_{i-1})^2 \right).$$

## A.2 Functional analysis

Among the many works on functional analysis, we refer the reader to [74, 79, 95, 122] and [125].

**Definition A.1** (**Normed vector space**). *Let $X$ be a vector space over the scalar field $\mathbb{R}$ (or $\mathbb{C}$). A norm (on $X$) is any mapping $\|\cdot\|_X : X \to \mathbb{R}$ which satisfies the following properties for all $x, y \in X$ and $c \in \mathbb{R}$ (or $\mathbb{C}$):*

*(i) positive homogenity: $\|cx\|_X = |c| \, \|x\|_X$,*

*(ii) subadditivity (or triangle inequality): $\|x + y\|_X \leq \|x\|_X + \|y\|_X$,*

*(iii) it separates points: $\|x\|_X = 0$ iff $x = 0$.*

*The pair $(X, \|\cdot\|_X)$ is called a normed vector space.*

Every normed vector space is a metric space with the metric $d_X : X \times X \to \mathbb{R}$ given by $d_X(x, y) = \|x - y\|_X$. The notions of open, closed, and closure of a set as well as the one of continuous operator can thus be introduced.

**Definition A.2** (**Strong convergence**). *Let $X$ be a normed vector space. A sequence $\{x_n\} \subset X$ converges to $x \in X$ ($x_n \to x$), iff*

$$\lim_{n\to\infty} \|x_n - x\|_X \equiv \lim_{n\to\infty} d_X(x_n, x) = 0.$$

*A sequence $\{x_n\} \subset X$ is called Cauchy iff for every real $\varepsilon > 0$ there exists a natural number $n \in \mathbb{N}$ such that for every integers $k, m > n$ it holds that $\|x_m - x_k\| < \varepsilon$.*

**Definition A.3** (**Banach space**). *A Banach space is a complete normed vector space, i.e. every Cauchy sequence is convergent.*

**Definition A.4** (**Inner product space**). *Let $X$ be a real (or complex) vector space. An inner product (on $X$) is any mapping $(\cdot, \cdot)_X : X \to \mathbb{R}$ (or $\mathbb{C}$) which satisfies the following properties for all $x, y \in X$ and $c \in \mathbb{R}$ (or $\mathbb{C}$):*

*(i) (conjugate) symmetry: $(x, y)_X = \overline{(y, x)}_X$,*

*(ii) linearity (in first argument): $(cx, y)_X = c \, (x, y)_X$,*

*(iii) positive definiteness: $(x, x)_X \geq 0$ with equality only for $x = 0$.*

*The pair $(X, (\cdot, \cdot)_X)$ is called an inner product space. The inner product $(\cdot, \cdot)_X$ at the same time induces the norm $\|\cdot\|_X$ which is given by $\|x\|_X = \sqrt{(x, x)_X}$ for $x \in X$.*

**Definition A.5** (**Hilbert space**). *A Hilbert space is an inner product space which is complete with respect to the induced norm.*

**Definition A.6 (Separability).** *A subset $M$ of a normed space $X$ is called dense iff its closure is the whole space $X$. A normed space $X$ is separable iff it contains a countable dense subset.*

**Definition A.7 (Compactness).** *A set $M$ in a normed vector space $X$ is called relatively compact iff every sequence in $M$ contains a convergent subsequence (i.e. having a limit in $X$). If $M$ is relatively compact and closed then it is called compact. Let $X, Y$ be normed vector spaces. An operator $A : X \to Y$ is called compact iff it maps bounded sets in $X$ on relatively compact sets in $Y$.*

**Definition A.8 (Imbeddings).** *Let $X, Y$ be Banach spaces with $\|\cdot\|_X, \|\cdot\|_Y$ and $X \subset Y$. We say that $X$ is continuously embedded in $Y (X \hookrightarrow Y)$ iff the identity map is continuous. We say that $X$ is compactly embedded in $Y$ $(X \hookrightarrow\hookrightarrow Y)$ iff the identity map is compact.*

**Definition A.9 (Linear bounded functional).** *A linear functional $f$ on a vector normed space $X$ is a linear mapping from $X$ to its field of scalars. A linear functional $f : X \to \mathbb{R}$ (or $\mathbb{C}$) is bounded iff the ratio $\frac{|f(x)|}{\|x\|_X}$ is bounded for all $x \in X, x \neq 0$.*

The action of the functional $f \in X^*$ on the vector $x \in X$ is often denoted as the dual pairing $\langle f, x\rangle$. The fact that every linear and bounded functional is continuous (i.e. $x_n \to x$ implies $f(x_n) \to f(x)$) leads to the following definition.

**Definition A.10 (Continuous dual space).** *A continuous dual space $X^*$ to the normed vector space $X$ is a normed vector space of all linear continuous functionals on $X$ with the operator norm*

$$\|f\|_{X^*} = \sup_{\|x\| \leq 1, x \neq 0} \frac{|\langle f, x\rangle|}{\|x\|}.$$

**Theorem A.1 (Hahn-Banach).** *Let $Y$ be a closed subspace of a Banach space $X$. If $f : Y \to \mathbb{R}$ is a linear bounded functional on $Y$, then there exists a linear functional $g : X \to \mathbb{R}$ which is the bounded extension of $f$ to the whole space $X$. That is to say that $g|_Y = f$ and $\|g\|_{X^*} = \|f\|_{Y^*}$.*

**Definition A.11 (Weak convergence).** *A sequence $\{x_n\}$ in the Banach space $X$ converges weakly to $x \in X$ $(x_n \rightharpoonup x)$ iff*

$$f(x_n) \to f(x) \qquad \text{for all } f \in X^*.$$

We also recall the notion of weak-$^*$ convergence:

$$f_n \overset{*}{\rightharpoonup} f \qquad \text{iff} \qquad f_n(x) \to f(x) \quad \text{for all } x \in X.$$

**Theorem A.2.** *The norm $\|\cdot\|_X$ in a Banach space $X$ is weakly sequentially lower semi-continuous on $X$, i.e*

$$\|x\|_X \leq \liminf_{n\to\infty} \|x_n\| \qquad \text{as } x_n \rightharpoonup x.$$

**Definition A.12 (Reflexivity).** *A normed space $X$ is reflexive iff it is isometrically isomorphic to its second dual $X^{**}$ by the canonical embedding*[1].

We note that in reflexive spaces both notions of the weak and weak-$^*$ convergence merge together.

The unit ball in an infinite dimensional Banach space is not compact what makes the analysis in infinite dimensions more difficult. The reflexivity property represents a partial remedy.

**Theorem A.3 (Weak compactness of reflexive spaces).** *The Banach space $X$ is reflexive if and only if every bounded sequence contains a weakly converging subsequence. In other words, the unit ball is weakly compact.*

**Theorem A.4 (Riesz representation theorem).** *For any continuous linear functional $f : X \to \mathbb{R}$ (or $\mathbb{C}$) on the Hilbert space $X$, there exists a unique element $y \in X$ such that*

$$f(x) \equiv \langle f, x \rangle = (x, y)_X \qquad \text{for all } x \in X,$$

*and moreover $\|f\|_{X^*} = \|y\|_X$.*

An immediate consequence of the Riesz representation theorem is that any Hilbert space is reflexive.

**Definition A.13 (Sesquilinear form).** *Let $X$ be a vector space over $\mathbb{K} = \mathbb{C}$. A mapping $a(\cdot, \cdot) : X \times X \to \mathbb{K}$ is called sesquilinear if it is linear in the first argument and conjugate-linear the second one, i.e. the conditions*

$$a(x + y, z + w) = a(x, z) + a(y, z) + a(x, w) + a(y, w),$$
$$\text{and} \quad a(tx, sy) = t\overline{s}\, a(x, y)$$

*hold for any $x, y, z \in X$ and $s, t \in \mathbb{K}$.*

*In the real case, if $\mathbb{K} = \mathbb{R}$, the mapping $a(\cdot, \cdot)$ is called bilinear.*

The sesquilinear form $a(\cdot, \cdot) : X \times X \to \mathbb{C}$ defines the dual mapping $A$

$$A : X \to X^*, \qquad u \mapsto a(u, \cdot).$$

**Theorem A.5 (Lax-Milgram).** *Let $A : X \to X^*$ be a linear continuous operator on the complex Hilbert space $X$. Suppose that there is a $c > 0$ such that*

$$|\langle Au, u \rangle| \geq c \|u\|^2 \quad \text{for all } u \in X. \tag{A.1}$$

*Then, for each given $b \in X^*$, the operator equation*

$$Au = b, \qquad u \in X.$$

*has a unique solution.*

[1] put more simple if $X = X^{**}$

Compare with [125, Theorem 18.E and 18.F] for the proof. Notice here the minor difference in the ellipticity assumption (A.1) to ensure the existence of a unique solution in the complex case.

## A.3 Nonlinear functional analysis

We begin with some notation. The spaces $X$ and $Y$ are always Banach spaces . The set $U(x) \subset X$ denotes an open neighborhood of $x$. For a map $r : U(0) \subseteq X \to Y$, we write

$$r(x) = o(\|x\|), x \to 0 \quad \text{iff } r(x)/\|x\| \to 0 \text{ as } x \to 0.$$

The class of all linear and continuous maps between Banach spaces $X$ and $Y$ is denoted by $L(X, Y)$.

**Definition A.14 (Variation).** *Let $f : U(x) \subseteq X \to Y$ be a given map. The variation (directional derivative) of $f$ at $x$ in the direction $h$ is defined as the limit*

$$\lim_{t \to 0} \frac{f(x + th) - f(x)}{t} = \delta f(x; h), \quad h \in X$$

*if it exists. If the mapping $h \mapsto \delta f(x; h)$ exists for all $h \in X$, it is called variation of $f$ at $x$.*

Note that $\delta f(x; h)$ need not be linear in $h$.

**Definition A.15 (Gâteaux and Fréchet derivative).** *Let $f : U(x) \subseteq X \to Y$ be a given map.*

*(1) The map $f$ is Fréchet differentiable at $x$ iff there exists a map $A \in L(X, Y)$ such that*

$$f(x + h) - f(x) = Ah + o(\|x\|), \qquad h \to 0 \tag{A.2}$$

*for all $h$ in some neighborhood of zero. If it exists, this $A$ is called Fréchet derivative of $f$ at $x$. We define $f'(x) = A$. The Fréchet differential at $x$ is defined as $df(x; h) = f'(x)h$.*

*(2) The map $f$ is Gâteaux differentiable at $x$ iff there exists a map $A \in L(X, Y)$ such that*

$$f(x + h) - f(x) = tAk + o(t), \qquad t \to 0 \tag{A.3}$$

*for all $k$ with $\|k\| = 1$ and all real numbers $t$ in some neighborhood of zero. $A$ is called Gâteaux derivative of $f$ at $x$. We define $f'(x) = A$. The Gâteaux differential at $x$ is defined as $df(x; h) = f'(x)h$.*

*(3) If the Fréchet (Gâteaux) derivatives exist for all $x \in U$, then the mapping*

$$f' : U \subseteq X \to L(X, Y), \qquad x \mapsto f'(x)$$

*is called Fréchet (Gâteaux) derivative of $f$ on $U$.*

*(4) Higher derivatives are defined successively. Thus, $f''(x)$ is the derivative of $f'$ at $x$. Higher differentials are defined as multilinear mappings.*

**Theorem A.6.** *The following holds*

*(a) If the variation $h \mapsto \delta f(x; h)$ is linear and continuous for all $h \in X$, then Gâteaux derivative of $f$ at $x$ exists and $f'(x)h = \delta f(x; h)$*

*(b) Gâteaux derivative at $x$ for which the passage to the limit in (A.3) is uniform for all $k$ with $\|k\| = 1$ is also a Fréchet derivative.*

*(c) If $f'$ exists as a Gâteaux derivative in some neighborhood of $x$ and $f'$ is continuous, then $f'(x)$ is also a Fréchet derivative.*

*(d) Every Fréchet derivative at $x$ is also a Gâteaux derivative.*

**Theorem A.7 (Hadamard's lemma).** *Suppose that $f : U \to Y$ is Fréchet differentiable, where $U$ be an open subset of $X$ If $x + \vartheta h \in U$ for any $\vartheta \in [0, 1]$, then we can write*

$$f(x + h) - f(x) = \int_0^1 f'(x + \vartheta h)h \, \mathrm{d}\vartheta,$$

*where the integral is understood in the Riemann sense.*

A mapping $f : U \subseteq X \to Y$, where $U$ is an open set, is called a $C^m-$map for $m \in \mathbb{N}$ iff $f$ has continuous Fréchet derivatives up to the order $m$ on $U$.

Let $M$ an $N$ be arbitrary sets in $X$ and $Y$. A mapping $f : M \to N$ is called a $C^m-$diffeomorphism iff $f$ is bijective and both $f$ and $f^{-1}$ are $C^m-$maps. A local $C^m-$diffeomorphism at $x_0$ is $C^m-$diffeomorphism from some neighborhood $U(x_0)$ in $X$ onto some neighborhood $U(f(x_0))$ in $Y$.

**Theorem A.8 (Inverse function theorem).** *Let $f : U(x_0) \subseteq X \to Y$ be a $C^1-$mapping Then $f$ is a local $C^1-$diffeomorphism at $x_0$ iff $f'(x_0)$ is bijective.*

The proof of the inverse function theorem is an application of the implicit function theorem, which itself relies on the Banach fixed-point theorem.

A mapping $f : X \to Y$ is called proper iff the preimage $f^{-1}(C)$ of every compact set $C$ in $Y$ is also compact.

**Theorem A.9 (Global inverse function theorem).** *Let $f : X \to Y$ be a local $C^m$-diffeomorphism, $0 \leq m \leq \infty$, at every point of $X$. Then $f$ is a $C^m$-diffeomorphism if and only if $f$ is proper.*

We refer reader to [124, Theorem 4.F and 4.G] for the proofs of the above statements and a comprehensive treatment of nonlinear functional analysis in general.

## A.4 Main theorem on monotone operators

The theory of monotone operators can be seen as a natural nonlinear extension of the ideas behind the (linear) Lax-Milgram theorem. A fundamental result on monotone operators was proved independently by Minty and Browder in 1963 (see the original articles [83] and [20]).

**Theorem A.10** (Main theorem on monotone operators)**.** *Let $X$ be a real reflexive Banach space. Assume that the operator $A : X \to X^*$ is*

*(i) monotone, i.e*

$$\langle Au - Av, u - v\rangle \geq 0 \qquad \text{for all } u, v \in X,$$

*(ii) hemicontinuous, i.e that the real function*

$$t \mapsto \langle A(u + tv), w\rangle$$

*is continuous on the interval $[0, 1]$ for all $u, v, w \in X$.*

*(iii) coercive, i.e*

$$\lim_{\|u\|_X \to \infty} \frac{\langle Au, u\rangle}{\|u\|_X} = +\infty.$$

*Then for each given $b \in X^*$, the operator equation*

$$Au = b \tag{A.4}$$

*has a solution $u \in X$. Provided that the operator $A$ is strictly monotone,*

$$\langle Au - Av, u - v\rangle > 0 \qquad \text{for all } u, v \in X \text{ with } u \neq v,$$

*the solution $u$ is unique.*

*Moreover, if the operator $A$ is strongly monotone, that is*

$$\langle Au - Av, u - v\rangle > c\,\|u - v\|_X^2 \qquad \text{for all } u, v \in X \text{ and fixed } c > 0,$$

*then $A^{-1}$ is Lipschitz continuous.*

Let us outline the proof of Theorem A.10 only for separable spaces. The whole proof can be found in [126, Section 26.2]. The basic idea of the proof is to replace the original operator equation (A.4) by finite-dimensional approximate equations and then prove the convergence of this approximation scheme. Such a technique is called the Galerkin method.

We first set

$$g(u) = \langle Au - b, u\rangle, \qquad g_i(u) = \langle Au - b, w_i\rangle.$$

The Galerkin system reads as follows

$$g_i(u_n) = 0, \quad i = 1 \dots n, \tag{A.5}$$

where

$$u_n \in X_n, \quad X_n = \text{span}\{w_1, \dots, w_n\} \quad \text{and} \quad u_n = \sum_{i=1}^{n} c_{in} w_i.$$

The union $\cup_{n=1}^{\infty} X_n$ is supposed to be dense in $X$.

It holds that a monotone and hemicontinuous operator on the real reflexive Banach space is demicontinuous, i.e.

$$u_n \to u \quad \text{implies} \quad Au_n \rightharpoonup Au.$$

For this reason the functions $g_i$ are continuous. The system (A.5) can be thus solved by means of Theorem (A.11) which itself follows from the Brouwer fixed-point theorem.

**Theorem A.11 (Existence principle).** *Let $B = \{x = (x_1, \dots, x_n) \in \mathbb{R}^n : |x| \le R\}$ for fixed $R > 0$. Let $g_i : B \to \mathbb{R}$ be continuous for $i = 1, \dots, n$. Assume that the condition*

$$\sum_{i=1}^{n} g_i(x) x_i \ge 0$$

*holds true on the set $|x| = R$. Then the system $g_i(x) = 0$, where $i = 1, \dots, n$, has a solution $x$ with $|x| \le R$.*

**Theorem A.12 (Brouwer fixed-point theorem).** *Suppose that $M$ is a nonempty, convex, compact set in $\mathbb{R}^n, n \ge 1$, and $f$ is a continuous mapping. Then $f$ has a fixed point, i.e there exists $x \in M$ such that $x = f(x)$.*

The coercivity assumption implies that the approximating sequence $u_n$ is bounded by a number $R > 0$

$$\|u_n\|_X \le R \quad \text{for all } n \in \mathbb{N}.$$

Since the monotone operators are locally bounded, it can be also derived that

$$\|Au_n\|_{X^*} \le C \quad \text{for all } n \in \mathbb{N}.$$

The existence of weakly convergent subsequences $u_n$ and $Au_n$ can be established by the reflexivity of $X$ and the density of $\cup_{n=1}^{\infty} X_n$, The monotonicity trick finally proves their convergence to the solution of the equation $Au = b$.

**Lemma A.1 (Monotonicity trick).** *Let $A : X \to X^*$ be a monotone and hemicontinuous operator on the real reflexive Banach space $X$. Assume that*

$$\begin{aligned} u_n &\rightharpoonup u \quad \text{in } X \qquad \text{as } n \to \infty, \\ Au_n &\rightharpoonup b \quad \text{in } X^* \qquad \text{as } n \to \infty, \\ \limsup_{n\to\infty} \langle Au_n, u_n\rangle &\le \langle b, u\rangle. \end{aligned}$$

*Then $Au = b$.*

The uniqueness for the strictly monotone operator $A$ follows from the uniqueness trick. In particular, if $Au = Av$ and $u \neq v$, then $\langle Au - Av, u - v\rangle > 0$ by the strict monotonicity which is a contradiction.

The strictly monotone operator $A$ is surjective and injective, therefore there exists the inverse operator $A^{-1}$. The Lipschitz continuity of $A^{-1}$ follows from the strong monotonicity of $A$.

## A.5 Variational calculus

We state the most important definitions of the variational calculus. The space $X$ always stands for a Banach space.

**Definition A.16 (Weak sequential lower semi-continuity).** *Let $f : M \supseteq X \to \mathbb{R}$. Then $f$ is called weakly sequentially lower semi-continuous iff for each $u \in M$ and each sequence $\{u_n\}$ in $M$,*

$$u_n \rightharpoonup u \quad \text{as } n \to \infty \qquad \text{implies} \qquad f(u) \le f(u_n).$$

**Definition A.17 (Convexity).** *A set $M \supseteq X$ is convex iff*

$$u, v \in M \quad \text{and} \quad t \in (0,1) \qquad \text{implies} \qquad tu + (1-t)v \in M.$$

*A functional $f : M \supseteq X \to \mathbb{R}$ is called convex iff*

$$f(tu + (1-t)v) \le tf(u) + (1-t)f(v)$$

*holds for any $u, v \in M$ and $t \in (0,1)$. The functional $f$ is called strictly convex iff the inequality is strict.*

*A functional $f : X \to \mathbb{R}$ is called proper convex iff it is convex and moreover iff $f(u) > -\infty, u \in X$ and $f \not\equiv \infty$.*

**Definition A.18 (Weak coercivity).** *Let $f : M \subseteq X \to \mathbb{R}$.*
*The functional $f$ is called coercive iff $f(u)/\|u\|_X \to \infty$ as $\|u\|_X \to \infty$ on $M$.*
*The functional $f$ is called weakly coercive iff $f(u) \to \infty$ as $\|u\|_X \to \infty$ on $M$.*

For the proofs of the next three statements see [126, Theorem 25.D, Proposition 25.20 and Theorem 25.F].

**Theorem A.13 (Main theorem on weakly coercive functionals).** *Suppose that the functional $f : M \subseteq X \to \mathbb{R}$ has the following three properties:*

*(i)* *$M$ is a nonempty convex set in the reflexive Banach space $X$.*

*(ii)* *$f$ is weakly sequentially lower semi-continuous.*

*(iii)* *$f$ is weakly coercive.*

*Then $f$ has a minimum on $M$.*

**Lemma A.2.** *Let $f : M \subseteq X \to \mathbb{R}$ be a functional on the convex closed set $M$ of the Banach space $X$. Then $f$ is weakly sequentially lower semi-continuous if one of the following three conditions is satisfied*

*(i)* *$f$ is continuous and convex.*

*(ii)* *$f$ is lower semicontinuous and convex.*

*(iii)* *$f$ is Gâteaux differentiable on $M$ and $df$ is monotone on $M$.*

**Theorem A.14 (Main theorem on monotone potential operators).** *Let $f : X \to \mathbb{R}$ be a Gâteaux differentiable functional on the reflexive Banach space $X$ with the following two properties:*

*(i)* *the Gâteaux derivative $df$ is monotone.*

*(ii)* *$f$ is weakly coercive.*

*Then:*

*(a)* *The minimum problem*

$$f(u) = \min!, \qquad u \in X,$$

*and the operator equation*

$$df(u) = 0, \qquad u \in X,$$

*are equivalent.*

*(b)* *Both problems have a solution.*

*(c)* *If $df$ is strictly monotone on $X$, then the solutions of the above problems are unique.*

## A.6 Function spaces

Our main reference here is the classical book [77]. The symbol $\Omega$ always denotes a domain with Lipschitz boundary, i.e. an open connected set in $\mathbb{R}^d, d > 1$ with the boundary $\partial\Omega$ which is locally described by the graph of a Lipschitz continuous function. We also adopt the following multi-index notation for partial derivatives:

$$D^\alpha u = \frac{D^{|\alpha|}u}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}},$$

where $\alpha = (a_1, \dots, a_n)$ is a multi-index of order $|\alpha| = k$.

**Theorem A.15 (Lebesgue dominated theorem).** *Let $\{f_n\}$ be a sequence of Lebesgue measurable functions $f_n : \Omega \to \mathbb{R}$. Assume that the sequence $\{f_n\}$ converges almost everywhere to Lebesgue measurable function $f$, and both are dominated by a function $g \in L^p(\Omega)$. Then both $f_n$ and $f$ belong to the space $L^p(\Omega)$ and moreover $f_n$ converges to $f$ in $L^p$-sense.*

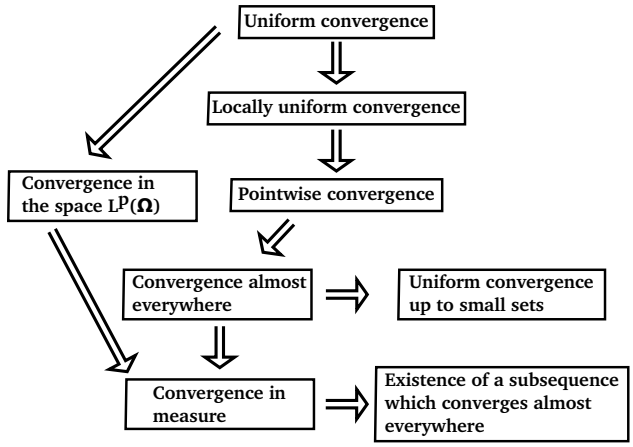



Figure A.1: An overview of relations between different types of convergences for an $f_n : \Omega \to \mathbb{R}$ (see [77, Fig. 2.12.1]).

The notion of Lebesgue integration and Lebesgue $L^p$ spaces are crucial to introduce Sobolev spaces on $\Omega$.

**Definition A.19 (Weak derivative).** *Let $u$ be a locally integrable function on $\Omega$. If there exists a locally integrable function $v$ such that the identity*

$$\int_\Omega u D^\alpha \varphi \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega v\varphi \, \mathrm{d}x$$

*holds for any smooth function $\varphi$ with a compact support on $\Omega$, then we call the $\alpha$-th weak derivative of $v$, $v = D^\alpha u$.*

**Definition A.20 (Sobolev spaces).** *The Sobolev space $W^{k,p}(\Omega)$ consists of all locally integrable functions $u : \Omega \to \mathbb{R}$ such that for every multi-index $\alpha$ with $|\alpha| < k$ the weak partial derivative belongs to $L^p(\Omega)$. We define the norm of $u \in W^{k,p}(\Omega)$ to be*

$$\|u\|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha|\le k} \|D^\alpha u\|_{L^p(\Omega)}\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \sum_{|\alpha|\le k} \operatorname{ess\,sup}_\Omega |D^\alpha u| & \text{if } p = \infty. \end{cases}$$

*If $p = 2$, we write sometimes $H^k(\Omega)$ instead of $W^{k,2}(\Omega)$.*

The space $W^{k,p}(\Omega)$ is reflexive if $1 < p < \infty$. It is separable if $1 \le p < \infty$. The following notation is sometimes used

$$|u|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{L^p(\Omega)}\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \sum_{|\alpha|=k} \operatorname{ess\,sup}_\Omega |D^\alpha u| & \text{if } p = \infty. \end{cases}$$

**Theorem A.16 (Sobolev imbeddings).** *Assume $d \ge 2$ and $1 \le p < \infty$. Then*

$$W^{1,p}(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$$

*provided either one of the following conditions is satisfied:*

*(i)* $p < d$, $1 \le q < dp/(d-p)$,

*(ii)* $p = d$, $q \in [1, \infty)$ *arbitrary.*

*If $q = dp/(d-p)$ and $d > p$, then $W^{1,p} \hookrightarrow L^q(\Omega)$.*

**Theorem A.17 (Trace theorem).** *There exists a uniquely determined continuous linear mapping $T : W^{1,p}(\Omega) \to L^q(\partial\Omega)$, where*

$$q = \frac{dp - p}{d - p} \quad \text{if } 1 \le p \le d \qquad \text{or} \qquad q \ge 1, \text{ arbitrary, if } d \le p,$$

*such that $Tu = u|_{\partial\Omega}$ for all smooth functions $u$ on $\Omega$.*

**Theorem A.18 (Equivalent norm by the Friedrichs inequality).** *Assume $\Gamma \subset \partial\Omega$ having a nonzero measure. Then there exists a constant $C$ depending only on $\Omega$ and $\Gamma$ such that*

$$\|u\|_{W^{1,2}(\Omega)} \le C \left(\|u\|^2_{L^2(\Gamma)} + \|\nabla u\|^2_{L^2(\Omega)}\right)^{1/2}$$

*for any $u \in W^{1,2}(\Omega)$.*

For the proof of the theorem above see [86]. The following theorem (see [77, Theorem 2.13.1]) can be thought as an $L^p$-version of the Arzela-Ascoli theorem.

**Theorem A.19 (Riesz-Kolmogorov theorem).** *Let $1 \leq p < \infty$. The set $K \subset L^p(\Omega)$ is relatively compact if and only if the two conditions are satisfied:*

*(i) The set $K$ is bounded, i.e. there exists a $C > 0$*

*(ii) The set $K$ is $p$-mean equicontinuous, i.e. for every $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$\int_\Omega |f(x+h) - f(x)|^p \,\mathrm{d}x < \varepsilon^p$$

*for each $f \in K$ and $h \in \mathbb{R}^d$ with $|h| < \delta$.*

**Lemma A.3 ([71, Lemma 1.3.10]).** *Let $X$ be a reflexive Banach space. Denote by $C((0,T),X)$ the space of all continuous function from the real interval $(0,T)$ to $X$ and by $C_w((0,T),X)$ the space of all $u : (0,T) \to X$ such that $\langle f, u(t)\rangle$ is continuous as a real function on $(0,T)$ for any $f \in X^*$.*

*(i) Let $u_n : (0,T) \to X, n \in \mathbb{N}$ be equibounded and equicontinuous. Then there exists $u \in C_w((0,T),X) \cap L^\infty((0,T),X)$ and a subsequence $\{u_{n_k}\}$ such that $u_{n_k}(t) \rightharpoonup u(t)$ in $X$ for all $t \in (0,T)$.*

*(ii) Let the imbedding $X \hookrightarrow Y$ be compact ($Y$ being a Banach space). If $u_n : (0,T) \to X, n \in \mathbb{N}$ is equibounded and $u_n : (0,T) \to Y, n \in \mathbb{N}$ is equicontinuous, then there exists $u \in C((0,T),Y) \cap L^\infty((0,T),X)$ and a subsequence $\{u_{n_k}\}$ such that $u_{n_k} \to u$ in $C((0,T),Y)$ and $u_{n_k} \rightharpoonup u(t)$ in $X$ for a.e. $t \in I$.*

## A.7 Finite element method

Let us recall a few basic notions of the finite element method. We closely follow the classical reference [31] by Ciarlet. A finite element in $\mathbb{R}^n$ is as a triple $(K, P_K, \Sigma_K)$ where:

(i) $K$ is a closed bounded set in $\mathbb{R}^n$ with nonempty interior and piecewise smooth boundary (*the element domain*),

(ii) $P_K$ is a finite dimensional space of functions over the set $K$ (*the space of basis functions*),

(iii) $\Sigma_K$ is a set of linearly independent functionals $\phi_j, j = 1, \ldots, N$ (*the degrees of freedom*).

By definition, the set $\Sigma_K$ is $P_K$-*unisolvent*, i.e. $\Sigma_K$ can be taken as a basis for the dual space $P'_K$. The functions $p_i \in P_K$ such that

$$\phi_j(p_i) = \delta_{ij} \qquad \text{for} \quad i, j = 1, \dots N$$

are called the *basis functions of the finite element*. Given a sufficiently smooth function $v : K \to \mathbb{R}$, so that degrees of freedom $\phi_j(v)$ for $j = 1, \dots N$ are well-defined, one can unambigously define the $P_K$-*interpolant* of $v$ as

$$\pi v = \sum_{j=1}^{N} \phi_j(v) p_i.$$

The set $K$ is usually a polyhedron in $\mathbb{R}^n$, but one can also consider "curved" elements. Most often, $K$ is an $n$-simplex (simplicial or triangular element domain) or $n$-rectangle. The space $P_K$ is (a subspace of) a *polynomial space* $P_k(K)$, with the polynomial degree less or equal to $k$.

**Example A.1** (**First-order Lagrange elements**). *The most used family of finite elements is Lagrange finite elements. Their degrees of freedom are point values. Let $\hat{K} \subset \mathbb{R}^n$ be the unit $n$-simplex,*

$$\hat{K} = \{(x_1, \dots x_n) \in \mathbb{R}^n : x_i \geq 0 \quad \forall i = 1, \dots, n \quad \textit{and} \quad \sum_{i=1}^{n} x_i \leq 1 \}.$$

*It is, in fact, the convex hull of the set $\{\boldsymbol{e}^j\}_{j=0}^n$, where $e_i^j = \delta_{ij}$. We define only the simplest member of this family, the first-order Lagrange finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$. The space $\hat{P}$ is the space of all the linear functions $P_1(\hat{K})$ and the corresponding set of degrees of freedom $\hat{\Sigma}$ is symbolically*

$$\hat{\Sigma} = \left\{ p(\boldsymbol{e^j}), \quad 0 \leq j \leq n \right\}.$$

**Example A.2** (**First-order edge elements**). *We recall the definition the first-order edge finite elements on tetrahedron. Let $\hat{K} \subset \mathbb{R}^3$ be the unit tetrahedron (the unit 3-simplex). The associated finite element space $\hat{P}$ consists of homogeneous linear vector polynomials $\boldsymbol{p(x)}$ such that $\boldsymbol{p} \cdot \boldsymbol{x} = \boldsymbol{0}$. Their degrees of freedom are line integrals along the edges of the tetrahedron*

$$\hat{\Sigma} = \left\{ \int_{\hat{e}} \boldsymbol{p} \cdot \boldsymbol{\tau} \, \mathrm{d}s, \quad \textit{for each edge } \hat{e} \textit{ of } \hat{K} \right\},$$

*where $\boldsymbol{\tau}$ is a unit vector in the direction of $\hat{e}$.*

**Remark A.1.** *The edge elements play an important role in electromagnetism. The reader is referred to [84] for a very well-written introduction into finite element method for Maxwell's equations. A clear and instructive introduction can be found for instance in [32]. We also wish to draw reader's attention to the the papers [8, 9, 7] by Arnold. They offer an unifying view on different families of finite element via the language of differential forms.*

We now move from "local" to "global" notions. Let $\Omega$ be a bounded domain with the boundary $\Gamma$ such that $\overline{\Omega} = \bigcup_{K \in T_h} K$. The set $T_h$ is called a mesh or *triangulation* of the domain. Here, it is essential to consider only *affine families of finite elements*, i.e., for which there exists a unique *invertible affine mapping* $F_K$ to the *reference element* $(\hat{K}, \hat{P}, \hat{\Sigma})$. This allows to describe the family $(K, P_K, \Sigma_K), K \in T_h$ as simply as possible. Rather than prescribing such a family by the data $K$, $P_K$ and $\Sigma_K$, it suffices to give one reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ and the affine mappings $F_K$, which is of practical and theoretical importance. The union of all the degrees of freedom $\Sigma_K$ gives a set of global degrees of freedom $\Sigma_h$. The associated *finite element space* $X_h$ consists of all "functions" $v$ such that $v|_K \in P_K, K \in T_h$ and $v$ fulfills some continuity conditions on the vertices (edges) of the adjacent elements [2]. The quotation mark indicates that $v$ need not be a proper function, since it need not have a unique definition along the faces common to adjacent finite elements. We say that $(K, P_K, \Sigma_K), K \in T_h$ is of *class* $C^0$, if the space $X_h$ is moreover a subset of the space of continuous functions on $\overline{\Omega}$ (compare with [31, Section 2.3]). Given a sufficiently smooth function $v : \overline{\Omega} \to \mathbb{R}$, the *(global)* $X_h$*-interpolant* of $v$ is defined as

$$\pi_h v = \sum_{j=1}^{M} \phi_{j,h}(v) w_j, \tag{A.6}$$

where $\phi_{j,h}$ are global degrees of freedom and $w_j$ are associated global basis functions.

We now briefly sketch a classical problem for the finite element method. Consider the variational problem

$$a(u, v) = f(v) \qquad \forall v \in V,$$

where the space $V$, the bilinear form $a$ and the linear form $f$ satisfy the assumptions of the Lax-Milgram lemma. To approximate the solution $u$, we first define the finite element space $V_h \subset X_h$. With each finite element space $V_h$ is associated the discrete solution $u_h$, which satisfies

$$a(u_h, v_h) = f(v_h) \qquad \forall v_h \in V_h.$$

The above equation is equivalent to the linear system

$$Ac = b, \qquad c \in \mathbb{R}^M,$$

[2] In case of the edge elements, for instance, one demands only the continuity of the tangential component of vector field across the edges of the adjacent elements.

where

$$A_{ij} = a(\phi_{i,h}, \phi_{j,h}), \qquad b_j = f(\phi_{j,h}),$$

and

$$u_h = \sum_{i=1}^{M} c_i w_i.$$

The standard choice of degrees of freedom ensures that the global basis functions are nonzero only on the adjacent elements. This leads to the sparse matrix $A$, which is easier to handle.

We finish this section with some convergence results.

A finite element method is *conforming* if $V_h \subset V$.

**Definition A.21 ([31, Section 3.2, (H1)]).** *A family of triangulations $T_h$ of the domain $\Omega$ is called regular iff*

*(i) There exists a constant $\sigma > 0$ such that*

$$h_K/\rho_k \leq \sigma \quad \text{for any simplex } K \in T_h,$$

*where $h_K$ is the diameter of $K$ and $\rho_k$ the supremum of the diameters of the spheres inscribed into $K$.*

*(ii) The discretization parameter $h$*

$$h = \max_{K \in T_h} h_K$$

*approaches zero.*

**Theorem A.20 ([31, Theorem 3.2.1]).** *Let $T_h$ be a regular family of triangulations of $\Omega$. Assume that all the finite elements $(K, P_K, \Sigma_K), K \in \bigcup_h T_h$ are affine-equivalent to a single reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ and they are of class $C^0$. Assume furthermore that there exist integers $k \geq 0$ and $l \geq 0$ with $l \leq m$, such that the following inclusions are satisfied:*

$$P_k(\hat{K}) \subset \hat{P} \subset H^l(\hat{K}),$$
$$H^{k+1}(\hat{K}) \hookrightarrow C^s(\hat{K}),$$

*where $s$ is the maximal order of partial derivatives occurring in the definitions of the set $\hat{\Sigma}$. Then there exists a constant $C$ independent of $h$ such that, for any function $H^{k+1}(\Omega)$,*

$$\|v - \pi_h v\|_{H^m(\Omega)} \leq C h^{k+l-m} |v|_{H^{k+1}(\Omega)}, \qquad 0 \leq m \leq \min\{1, l\}, \tag{A.7}$$

*where $\pi_h v$ is the $X_h$-interpolant of the function $v$.*

# Bibliography


[1] R. Acar and C. R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10(6):1217–1229, 1994.

[2] G. Allaire. *Shape Optimization by the Homogenization Method*, volume 146 of *Applied Mathematical Sciences*. Springer, 2002.

[3] G. Allaire, F. de Gournay, F. Jouve, and A.-M. Toader. Structural optimization using topological and shape sensitivity via a level set method. *Control and Cybernetics*, 34(1):59–80, 2005.

[4] G. Allaire, F. Jouve, and A.-M. Toader. Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, 194(1):363–393, 2004.

[5] Eugene L. Allgower and Kurt Georg. *Introduction to Numerical Continuation Methods*. SIAM, 2003.

[6] H.W. Alt and S. Luckhaus. *Quasilinear elliptic-parabolic differential equations*, volume 3 of *183*. MATHEMATISCHE ZEITSCHRIFT, 1983.

[7] D. N. Arnold and H. Chen. Finite element exterior calculus for parabolic problems. *submitted to SIAM Journal on Numerical Analysis*.

[8] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006.

[9] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus: from Hodge theory to numerical stability. *Bull. Amer. Math. Soc. (N.S.)*, 47:281–354, 2010.

[10] K. Astala and L. Paivarinta. Calderone's inverse conductivity problem in the plane. *Annals of Mathematics*, 163:265–299, 2006.

[11] D. Bainov and P. Simeonov. *Integral inequalities and applications*, volume 57. Springer, 1992.

[12] C. Baiocchi and A. Capelo. *Variational and Quasivariational Inequalities Application to Free Boundary Problems.* Wiley, 1984.

[13] G. Barbero and L. Pandolfi. Surface viscosity in nematic liquid crystals. *Physical Review E*, 79(5):051701, 2009.

[14] J. Bear. *Dynamics of Fluids in Porous Media*. Dover Publ., 1972.

[15] M.P. Bendsøe and N. Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, 71(2):197–224, 1988.

[16] J. P. Berenger. A perfectly matched layer for the absorption of electromagnetic wave. *J. Comput. Phys.*, 114:185–200, 1994.

[17] L. Borcea. Electrical impedance tomography. *Inverse Problems*, 18(6), 2002.

[18] A. Bossavit. *Computational electromagnetism. Variational formulations, complementarity, edge elements*, volume XVIII of *Electromagnetism*. Academic Press, Orlando, FL., 1998.

[19] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer, 2010.

[20] F.E. Browder. Nonlinear elliptic boundary value problems. *Bull. Amer. Math. Soc.*, 69(6), 1963.

[21] P. Brunner, R. Merwa1, A. Missner, J. Rosell, K. Hollaus, and H. Scharfetter. Reconstruction of the shape of conductivity spectra using differential multifrequency magnetic induction tomography. *Physiological Measurement*, 27:237–248, 2006.

[22] A. Buffa, M. Costabel, and D. Sheen. On traces for $\boldsymbol{H}(\mathrm{curl};\Omega)$ in Lipschitz domains. *Journal of mathematical analysis and applications*, 276(2):845–867, 2002.

[23] M. Burger, B. Hackl, and W. Ring. Incorporating topological derivatives into level set methods. *Journal of Computational Physics*, 194(1):344–362, 2004.

[24] A. Calderon. On an inverse boundary value problem. *Computational and applied mathematics*, 25:133–138, 2006.

[25] C. Cavaterra, M. Gal, C. G. Grasselli, and Miranville A. Phase-field systems with nonlinear coupling and dynamic boundary conditions. *Nonlinear Analysis*, 72:2375–2399, 2010.

[26] J. Cea and K. Malanowski. An example of a max-min problem in partial differential equations. *SIAM Journal on Control*, 8(3):305–316, 1970.

[27] M. Cessenat. *Mathematical methods in electromagnetism: Linear theory and applications*. World Scientific, 1996.

[28] Tony F Chan and Xue-Cheng Tai. Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients. *Journal of Computational Physics*, 193(1):40–66, 2004.

[29] M. Cheney, D. Isaacson, and J.C. Newell. Electrical impedance tomography. *SIAM Review*, 41(1):85–101, 1999.

[30] ET Chung, TF Chan, and XC Tai. Electrical impedance tomography using level set representation and total variational regularization. *JOURNAL OF COMPUTATIONAL PHYSICS*, 205(1):357–372, 2005.

[31] Phillipe G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2002.

[32] I. Cimrák. *On Landau-Lifshitz equation of ferromagnatism.* PhD thesis, Ghent University, 2005.

[33] I. Cimrák and V. Melicher. Determination of precession and dissipation parameters in micromagnetism. *Journal of Computational and Applied Mathematics*, 234(7):2239–2249, 2010.

[34] G. M. Coclite, A. Favini, , G.R. Goldstein, J.A Goldstein, and Romanelli S. Continuous dependence on the boundary conditions for the wentzell laplacian. *Semigroup Forum*, 77:101–108, 2008.

[35] G.M Coclite, G.R. Goldstein, and J.A Goldstein. Stability of parabolic problems with nonlinear Wentzell boundary conditions. *Journal of Differential Equations*, 246:2434–2447, 2009.

[36] David Colton and Rainer Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93. Springer, 2013.

[37] De Cezaro, A. and Leitao, A. and Tai, X.-C. On piecewise constant level-set (PCLS) methods for the identification of discontinuous parameters in ill-posed problems. *Inverse Problems*, 29, 2013.

[38] M. Duruflé, H. Haddar, and P. Joly. Higher order generalized impedance boundary conditions in electromagnetic scattering problems. *Comptes Rendus Physique*, 7(5):533–542, 2006.

[39] C. Ebmeyer. Error estimates for a class of degenerate parabolic equations. *SIAM journal on numerical analysis*, 35(3):1095–1112, 1998.

[40] M. Eller, J. E. Lagnese, and S. Nicaise. Decay rates for solutions of a Maxwell system with non linear boundary damping. *Computational and Applied Mathematics*, 21(1):135–165, 2002.

[41] M. Eller, J. E. Lagnese, and S. Nicaise. Stabilization of heterogeneous Maxwell's equations by linear or nonlinear boundary feedbacks. *EJDE*, 2002(21):1–26, 2002.

[42] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

[43] B. Engquist and A. Majda. Absorbing boundary conditions for numerical simulation of waves. *Math. Comp.*, 31(139):629–651, 1977.

[44] H. A. Eschenauer, V. V. Kobelev, and A. Schumacher. Bubble method for topology and shape optimization of structures. *Structural and Multidisciplinary Optimization*, 8:42–51, 1994.

[45] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, 1998.

[46] M. Fabrizio and A. Morro. *Electromagnetism of continuous media. Mathematical modelling and applications*. Oxford University Press, Oxford, 2003.

[47] W. Fang and K. Ito. Identification of contact regions in semiconductor transistors by level-set methods. *Journal of Computational and Applied Mathematics*, 159(2):399–410, 2003.

[48] A. Favini, G. R. Goldstein, J. A. Goldstein, and S. Romanelli. $C_0$-semigroups generated by second order differential operators with general Wentzell boundary condition. *Proc. Amer. Math Soc.*, 128:1982–1989, 2000.

[49] M. Fecko. *Differential Geometry and Lie Groups for Physicists*. Cambridge University Press, 2006.

[50] P. Fulmanski, A. Laurain, J.-F. Scheid, and J. Sokolowski. Level set method with topological derivatives in shape optimization. *Int. J. Comput. Math.*, 85:1491–1514, 2008.

[51] C. G. Gal. Well-posedness and long time behavior of the non-isothermal viscous Cahn-Hilliard equation with dynamic boundary conditions. *Dynamics of PDE*, 5(1):39–67, 2008.

[52] C. G. Gal. On a class of degenerate parabolic equations with dynamic boundary conditions. *Journal of Differential Equations*, 253:126–166, 2012.

[53] C. Geuzaine, P. Dular, and W. Legros. Dual formulations for the modeling of thin electromagnetic shells using edge elements. *IEEE Transactions on Magnetics*, 36:799–803, 2000.

[54] G. R. Goldstein. Derivation and physical interpretation of general boundary conditions. *Adv. Differential Equations*, 11:457–480, 2006.

[55] H. Griffiths. Magnetic induction tomography. *Measurement Science and Technology*, 12:1126–1131, 2001.

[56] D. Gursoy and H. Scharfetter. Imaging artifacts in magnetic induction tomography caused by the structural incorrectness of the sensor model. *Measurement Science and Technology*, 22, 2011.

[57] J. Gyselinck, R.V. Sabariego, P. Dular, and C. Geuzaine. Time-domain finite-element modeling of thin electromagnetic shells. *Magnetics, IEEE Transactions on*, 44(6):742–745, 2008.

[58] Barucq H. A new family of first-order boundary conditions for the maxwell system: derivation, well-posedness and long-time behavior. *J. Math. Pures Appl.*, 82:67–88, 2002.

[59] H. Haddar and P. Joly. Stability of thin layer approximation of electromagnetic waves scattering by linear and nonlinear coatings. *Journal of Computational and Applied Mathematics*, 143:201–236, 2002.

[60] Lin He, Chiu-Yen Kao, and Stanley Osher. Incorporating topological derivatives into shape derivatives based level set methods. *Journal of Computational Physics*, 225(1):891–909, 2007.

[61] Frédéric Hecht, Olivier Pironneau, Jacques Morice, Antoine Le Hyaric, and Kohji Ohtsuka. *FreeFem++*. Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, 3rd edition, May 2009.

[62] M. Hintermueller and A. Laurain. Electrical impedance tomography: from topology to shape. *Control and Cybernetics*, 37(4, SI):913–933, 2008.

[63] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica 11.0*, pages 237–339, 2002.

[64] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, 2007.

[65] N. Ida and S. Yuferev. Impedance boundary conditions for transient scattering problems. *IEEE Transactions on Magnetics*, 33(3):1444–1447, 1997.

[66] V. Isakov. *Inverse problems for partial differential equations*, volume 127. Springer Verlag, 1998.

[67] J. D. Jackson. *Classical Electrodynamics*. John Wiley and Sons, 1999.

[68] J. Jost. *Riemannian Geometry and Geometric Analysis*. Springer, 2005.

[69] J. Kačur. Nonlinear parabolic boundary value problems with the time derivative in the boundary conditions. In *Equadiff IV*, volume 703 of *Lecture Notes in Mathematics*, pages 170–178, 1979.

[70] J. Kačur. Nonlinear parabolic equations with the mixed nonlinear and nonstationary boundary conditions. *Mathematica Slovaca*, 30(3):213–237, 1980.

[71] J. Kačur. *Method of Rothe in evolution equations*, volume 80 of *Teubner Texte zur Mathematik*. Teubner, Leipzig, 1985.

[72] J. Kačur. Solution to strongly nonlinear parabolic problems by a linear approximation scheme. *IMA J. Numer. Anal.*, 19(1):119–145, 1999.

[73] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Springer, 2011.

[74] A. Kolmogorov and S. Fomin. *Elements of the theory of functions and functional analysis*, volume 1. Dover Publications, 1999.

[75] A. Korjenevsky, V. Cherepin, and S. Sapetsky. Magnetic induction tomography: experimental realization. *Physiological Measurement*, 21:89–94, 2000.

[76] L. Krahenbuhl and D. Muller. Thin layers in electrical engineering-example of shell models in analysing eddy-currents by boundary and finite element methods. *IEEE Transactions on Magnetics*, 29(2):1450–1455, 1993.

[77] A. Kufner, O. John, and S. Fučík. *Function Spaces*. Monographs and textbooks on mechanics of solids and fluids. Noordhoff International Publishing, Leyden, 1977.

[78] A. Logg, K.-A. Mardal, G. N. Wells, and et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[79] B. MacCluer. *Elementary functional analysis*, volume 253. Springer, 2008.

[80] James Clerk Maxwell. A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155(459512), 1865.

[81] I. Mayergoyz. *Nonlinear Diffusion of Electromagnetic Fields: With Applications to Eddy Currents and Superconductivity*. Academic Press, 1998.

[82] A. Megrabov. *Forward and inverse problems for hyperbolic, elliptic, and mixed type equations*. Number 40. Vsp, 2003.

[83] G.J. Minty. On a monotonicity method for the solution of nonlinear equation in Banach spaces. *Proc. Nat. Acad. Sci. USA*, 50(6), 1963.

[84] P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2003.

[85] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Berlin: Springer, 1984.

[86] J. Nečas. *Les méthodes directes en théorie des équations elliptiques*. Academia, Prague, 1967.

[87] L. K. Nielsen, X.-C. Tai, Si. I. Aanonsen, and M. Espedal. A binary level set model for elliptic inverse problems with discontinuous coefficients. *INTERNATIONAL JOURNAL OF NUMERICAL ANALYSIS AND MODELING*, 4(1):74–99, 2007.

[88] R. Nochetto and C. Verdi. Approximation of degenerate parabolic problems using numerical integration. *SIAM Journal on Numerical Analysis*, 25(4):784–814, 1988.

[89] S. Osher and J.A. Sethian. Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79:12–49, 1988.

[90] O. Pironneau. *Optimal Shape Design for Elliptic Systems*. Springer Netherlands (Springer-Verlag, Berlin, Heidelberg, New York, Tokyo), 1984.

[91] C. Poignard. Asymptotics for steady-state voltage potentials in a bidimensional highly contrasted medium with thin layer. *Mathematical Methods in the Applied Sciences*, 31(4):443–479, 2008.

[92] C. Poignard and et al. Approximate conditions replacing thin layers. *IEEE TRANSACTIONS ON MAGNETICS*, 44(6):1154–1157, 2008.

[93] F. Radu, I. Pop, and P. Knabner. Error estimates for a mixed finite element discretization of some degenerate parabolic equations. *Numerische Mathematik*, 109(2):285–311, 2008.

[94] K. Rektorys. *Variational methods in mathematics, science, and engineering*. D. Reidel Publishing Company, 1980.

[95] W. Rudin. Functional analysis. International series in pure and applied mathematics, 1991.

[96] P. Russer. *Electromagnetics, microwave circuit and antenna design for communications engineering*. Artech House Publishers, 2003.

[97] F. Santosa. A level-set approach for inverse problems involving obstacles. *ESAIM Contrôle Optim. Calc. Var.*, 1:17–33, 1996.

[98] A. Schumacher. *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungkriterien*. PhD thesis, Siegen University, Siegen, Germany, 1996.

[99] T.B.A. Senior and J.L. Volakis. *Approximate Boundary Conditions in Electromagnetics*. IET, 1995.

[100] M. Slodička. Semigroup formulation of rothes method: Application to parabolic problems. *CMUC*, 33:245–260, 1992.

[101] M. Slodička. A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media. *J. Sci. Comput.*, 23(5):1593–1614, 2002.

[102] M. Slodička. An approximation scheme for a nonlinear degenerate parabolic equation with a second order differential Volterra operator. *JCAM*, 168(447–458), 2004.

[103] M. Slodička and V. Zemanová. Time-discretization scheme for quasi-static Maxwell's equations with a non-linear boundary condition. *J. Comput. Appl. Math.*, 216(2):514–522, 2008.

[104] J. Sokołowski and A. Zochowski. On topological derivative in shape optimisation. Technical Report 3170, INRIA-Lorraine, 1997.

[105] J. Sokołowski and A. Zochowski. On the topological derivative in shape optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, 1999.

[106] J. Sokołowski and A. Zochowski. Topological derivatives for elliptic problems. *Inverse Problems*, 15(1):123–134, 1999.

[107] J. Sokołowski and J. P. Zolésio. *Introduction to shape optimization. Shape sensitivity analysis.*, volume 16. Springer-Verlag, 1992.

[108] M. Soleimani. Computational aspects of low frequency electrical and electromagnetic tomography: a review study. *International Journal For Numerical Analysis and Modeling*, 5(3):407–440, 2008.

[109] N. Su. Multidimensional degenerate diffusion problem with evolutionary bopundary condition: Existence, uniqueness and approximation. *International Series of Numerical Mathematics*, 114:165–178, 1993.

[110] K. Sungwhan and Y. Masahiro. Uniqueness in the two-dimensional inverse conductivity problems of determining convex polygonal supports: case of variable conductivity. *Inverse Problems*, 20:495–506, 2004.

[111] X.-C. Tai and H. Li. A piecewise constant level set method for elliptic inverse problems. *APPLIED NUMERICAL MATHEMATICS*, 57(5-7):686–696, 2007.

[112] K. Tarumi and H. Schwegler. A nonlinear treatment of the protocell model by a boundary layer approximation. *Bulletin of Mathematical Biology*, 49(3):307–320, 1987.

[113] A. N. Tikhonov. The stability of inverse problems. *Doklady Akad. Nauk SSSR*, 39(5):195–198, 1943. in Russian.

[114] A. N. Tikhonov. Regularization of incorrectly posed problems. *Doklady Akad. Nauk SSSR*, 153(1):49–52, 1963. in Russian.

[115] M.M. Vajnberg. *Variational method and method of monotone operators in the theory of nonlinear equations.* John Wiley & Sons, 1973.

[116] J.L. Vazquez and Vitillaro E. Heat equation with dynamical boundary conditions of reactive-diffusive type. *Journal of Differential Equations*, 250(4):2143–2161, 2011.

[117] V. Vrábeľ and M. Slodička. An eddy current problem with a nonlinear evolution boundary condition. *Journal of Mathematical Analysis and Applications*, 387(1):267–283, 2012.

[118] V. Vrábeľ and M. Slodička. Nonlinear diffusion problem with dynamical boundary value. *Journal of Computational and Applied Mathematics*, 246:94–103, 2012.

[119] P. Šolín. Partial differential equations and the finite element method. *Trans. Magn*, 18(43):1435, 1982.

[120] K.F. Warnick, R.H. Selfridge, and D.V. Arnold. Teaching electromagnetic field theory using differential forms. *Education, IEEE Transactions on*, 40(1):53–68, 1997.

[121] A. D. Wentzell. On boundary conditions for multi-dimensional diffusion processes. *Theor. Probability Appl.*, 4:164–177, 1959.

[122] K. Yosida. *Functional analysis*, volume 35. Springer, Berlin, 1980.

[123] S. V. Yuferev and N. Ida. *Surface Impedance Boundary Conditions: A Comprehensive Approach*. CRC PressINC, 2009.

[124] E. Zeidler. *Nonlinear Functional Analysis and its Applications I:Fixed-Point Theorems*. Springer-Verlag, 1986.

[125] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/A: Linear Monotone Operators*. Springer-Verlag, 1990.

[126] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*. Springer-Verlag, 1990.

[127] V. Zemanová. *On numerical methods for direct and inverse problems in electromagnetism.* PhD thesis, Ghent University, 2009.

[128] S. Zhu, Q. Wu, and C. Liu. Shape and topology optimization for elliptic boundary value problems using a piecewise constant level set method. *Applied Numerical Mathematics*, 61(6):752–767, 2011.

[129] M. Zolgharni, P. D. Ledger, and Griffiths H. Forward modelling of magnetic induction tomography: a sensitivity study for detecting haemorrhagic cerebral stroke. *Medical and Biological Engineering and Computing*, 47(12):1301–1313, 2009.